DOCUMENT RESUME

ED 131 094 TM 005 787

TITLE	Aspects of Educational Assessment.
INSTITUTION	Educational Testing Service, Princeton, N.J. Center for Statewide Educational Assessment.
SPONS AGENCY	Ford Foundation, New York, N.Y.
PUB DATE	75
NOTE	166p.; For the individual papers included here, see ED 074 071-072, ED 080 533-534, ED 093 990 and ED 097 376
AVAILABLE FROM	Educational Testing Service, Princeton, N.J. 08540 ($4.95)
EDRS PRICE	MF-$0.83 HC-$8.69 Plus Postage.
DESCRIPTORS	Academic Achievement; *Attitude Tests; Data Analysis; Data Collection; Definitions; *Educational Assessment; Educational Status Comparison; *Evaluation Methods; Guides; *Sampling; School Attitudes; Self Concept; *Self Concept Tests; *State Programs; State Surveys; Student Testing; *Test Construction; Testing Programs; Test Interpretation; Test Reviews; Test Selection

ABSTRACT

Six research papers that have been published by the staff of the Center for Statewide Educational Assessment at Educational Testing Service are presented. In "A Selection of Self Concept Measures," Joan Knapp explores questions of defining and measuring self concept. In another paper, she examines some problems of measuring attitudes toward school. In both papers she comments upon the use and value of a number of instruments. Richard Jaeger's paper, "A Primer on Sampling for Statewide Assessment," is designed to help the reader meet a sampling expert at least half way. In "The Use of Correlates of Achievement in Statewide Assessment," Paul Campbell outlines testing strategies that take into account the relationship between learning conditions and achievement. John Fremer suggests ways to determine what should be measured, whether newly developed or existing instruments should be used, and what types of reports are needed in his paper, "Developing Tests for Assessment Programs: Issues and Suggested Procedures." In the final paper of this volume, "Statewide Assessment: Methods and Concerns," Nancy Bruno, Paul Campbell, and William Schabacker present a comprehensive guide for state assessment personnel. Their step-by-step approach provides answers to these questions: How do we involve the community in testing programs? How can data presentations be designed for laymen? How can we take the noncognitive effects of school into account? (RC)

Aspects of Educational Assessment

CENTER FOR STATEWIDE EDUCATIONAL ASSESSMENT
EDUCATIONAL TESTING SERVICE • PRINCETON, NEW JERSEY

ASPECTS OF EDUCATIONAL ASSESSMENT

Published by the Center for Statewide
Educational Assessment, which is supported
by funds from the Ford Foundation

CENTER FOR STATEWIDE EDUCATIONAL ASSESSMENT
EDUCATIONAL TESTING SERVICE - PRINCETON, NEW JERSEY

CONTENTS

FOREWORD

The Center for Statewide Educational Assessment was established under a grant from the Ford Foundation to provide consulting, instructional, and information services for the development of statewide educational assessment programs. The primary focus of the Center has been on improving the capabilities of assessment personnel in state departments of education.

We are pleased to present here a number of research papers that have been published by the Center's staff. In these papers, the authors come to grips with some of the most difficult problems of measurement. Joan Knapp explores questions of defining and measuring self concept: Which of the many overlapping definitions of self concept should form the basis for measurement? Would it be better to think of self concept as a field of study rather than a trait? In another paper, Joan Knapp examines some problems of measuring attitudes toward school. In both papers, the author comments upon the use and value of a number of instruments. Richard Jaeger's "primer" on sampling, designed to help the reader "meet a sampling expert at least half way," is a valuable resource presented with humor and imagination. Paul Campbell outlines testing strategies that take into account the relationship between learning conditions and achievement. John Fremer suggests ways to determine what should be measured, whether newly developed or existing instruments should be used, and what types of reports are needed. In the final paper of this volume, Nancy Bruno, Paul Campbell, and William Schabacker present a comprehensive guide for state assessment personnel. Their step-by-step approach provides answers to some challenging questions: How do we involve the community in testing programs? How can data presentations be designed for laymen? How can we take the noncognitive effects of school into account?

We are grateful for the opportunity to offer this research which, we hope, will expand the horizons of statewide educational assessment.

William W. Turnbull
President
Educational Testing Service

A SELECTION OF SELF CONCEPT MEASURES

Joan Knapp

TABLE OF CONTENTS

Introduction

The study of the self is a fairly recent development in the history of psychology. The work and theories of Freud (although he never used the term 'self') and the writings of William James promoted some interest in the topic in psychological circles. Unfortunately, the theoretical foundation for studies concerning the self and self concept was not completely laid before behaviorism emerged and dominated psychological thought for the first four decades of this century. Wylie (1961) points out that when American clinical psychologists discovered that stimulus-response models were too limited to be applied to therapeutic settings, interest in the self and self concept was renewed and great energy was directed toward research activity in this area. More recently, the desire to enhance the self concepts of children as students, particularly in early childhood education, and the logical connection between self concept and achievement have stimulated educational studies and assessment in this area.

Definition of Self Concept

Because of this historical unevenness in the development of theories concerning self concept, a study of the literature and the state of the art reveals an endless list of terms such as social self, self regard, self esteem, self evaluation, phenomenal self, self image, etc. Many of these terms have overlapping definitions, and the theories associated with them are ambiguous and incomplete with no one theory receiving a large amount of meticulous empirical exploration. Thus, when the evaluator's or educator's task is to study self concept in the school setting, he is faced with the dilemma of not knowing exactly what he is studying and, of course, how he is to assess or measure its extremes or changes.

Because of this confusion, it may be wise at this point to think of self concept as a term that designates a field of study rather than a unified construct or trait. It is a term given to a set of self referent constructs which form a unique collection of complex and dynamic ideas. A person may or may not be aware of the ideas he/she holds true about him/herself in respect to a given situation, however we can assume that a person's self concept or an aspect of its affects his/her behavior (Coller, 1971). Self concept defined as a multidimensional construct that covers and includes the total range of one's perceptions and evaluations of oneself (Creelman, 1954) is a widely acknowledged and less technical definition.

The Measurement of Self Concept

It is obvious that as much as we would like to put ourselves in someone else's skin, it cannot be done. We cannot feel or see a person's self concept; therefore, it must be inferred by using various measurement techniques. Coller (1971) has offered a useful model (adapted from Gordon, 1968) that provides gross but useful categories for the classification of measurement devices (see figure on page 3).

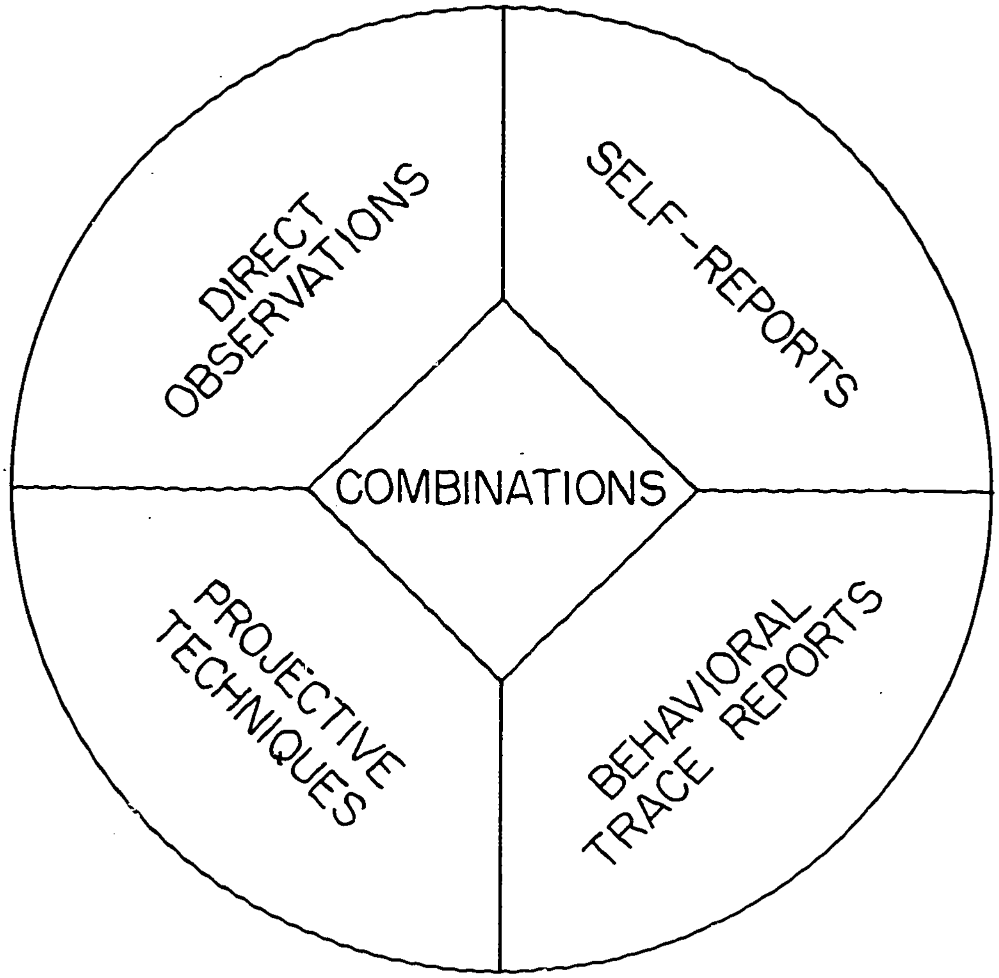
Each type of measure has methodological flaws and advantages. Direct observations are useful for very young children who cannot use language with facility and who have attention spans too short for a testing situation. However, the presence of the observer may produce behavior on the part of the subject which is different than the subject's behavior would be if the observer were not present.

Behavior trace measures eliminate this observer effect as the student is unaware that his behavior is being studied. These procedures are concerned with examining the aftereffect produced by a child's responses, not with direct observation. Trace behavior techniques may entail such things as studying comments in a student's school record files or evaluating in retrospect a child's self concept on a rating scale by way of impressions of a child's behavior in the classroom. However, since the investigator is never sure what behavior is reflected by file comments and since memories may be faulty or distorted, the data obtained may be inaccurate.

Projective techniques which use unstructured test stilumi such as inkblots or pictures are effective in revealing latent and covert aspects of self concepts, are less likely to be subject to faking, and are useful with verbally limited individuals. But scoring is difficult and may lack objectivity. Interpretation of scores can result in a misleading picture of the subject, and the determination of reliability and validity present special problems.

Self report techniques are economical and practical in that they can be scored and interpreted easily, and the investigator can obtain a self description from a subject in a short period of time because the measures are structured or semistructured. On the minus side, there is evidence that subjects can recognize items or answers on instruments, such as questionnaires, which are socially more desirable than others and therefore can "fake good" or "fake bad" depending on the circumstances surrounding the self report. However, much of this can be eliminated by taking this into account when the instrument

A General Model for the Assessment of Self



The circle represents all that is meant by Self and includes all definitions. The diamond shape in the center represents Self as assessed by any combination of four distinct procedures: Direct Observation, Behavioral Traces, Self-Reports, and Projective Techniques.

is constructed (e.g., using equal numbers of negative and positive statements), by establishing rapport with the student, by providing a nonthreatening climate, and by assuring anonymity when administering the self report. The majority of self concept measures used in research consists of self report inventories.

Clearly, since each type of measure has weaknesses, any assessment of self concept should employ an eclectic approach. In research and evaluation, an investigator can be more confident in the results of his assessment when several different measurement methods produce comparable findings.

Caveat Emptor

Before undertaking large-scale assessments in the area of student self concepts, the educator, researcher, and evaluator should be aware of the pitfalls, problems, and eddies of confusion which abound concerning the topic in the disciplines of psychology, sociology, and education.

The major problem, and one from which most other problems stem, was touched upon earlier--the lack of cohesiveness and tight conceptualization concerning self and self concept, and yet this can be said of many areas studied in the social sciences. Since it is clear that this problem will not be remedied quickly, investigators can contribute to a solution by prefacing and supporting their assessment procedures with a clear and precise rationale. That is, self concept should be described theoretically as well as operationally. Frequently, reports of self concept research do not even provide a good description of the instrument used and/or the reasons for its use.

There are problems concerning the psychometric properties of the instruments. Personality or noncognitive measures generally are less stable than cognitive measures; yet many instruments in the field are substantiated with internal consistency coefficients when test-retest reliability data would be more meaningful and appropriate. In terms of validity, instrument developers and users have relied heavily on expert judgement and theories which may lead to content validation, but which do not speak to construct or criterion related validation. Very few instruments have undergone convergent and discriminant validation—that is, the study of the interrelationships between more than one method of measuring self concept and other constructs which may be similar or dissimilar to self concept. Construct validation is assured if different measures of the same trait or construct correlate higher with each other than they do with measures of different traits involving separate methods (Campbell and Fiske, 1959). More simply, the caution here is to take more than one measurement approach when planning self concept assessment.

There are other unanswered questions and unresolved issues which may influence the design of research and evaluation in this area. A few are listed below.

1. Does low self concept result in poor achievement, or does poor achievement result in a lowered self concept?
2. How much do response sets and defensiveness on the part of subjects affect their scores on a self concept measure--in particular, self reports?
3. How stable is self concept at different ages in a child's life?
4. Can self concept be changed? If so, what procedures or teaching styles work?
5. Is self concept differentiated or global?
6. Does sex role identification influence self concept?
7. Do minority group children have lower self esteem than majority group children? All the time? Under certain conditions?
8. Do particular cultures influence the way individuals evaluate themselves?

Instruments

The following instruments, as a group, have been chosen on the basis of several criteria.

1. They should be suitable for and reflect the full age range of children in school.
2. Each of the categories in Coller's model--self report, projective, behavior trace, and direct observation--should be represented.
3. They should have been designed with the so-called normal population in mind rather than a psychopathological population.
4. They have enough information accompanying them to enable investigators to use them effectively.
5. They should reflect a variety of means of presentation (e.g., pictorial items, semantic differential).

Direct Observation

Title: Work Posting*

Description: This measure is one of a collection of instruments concerning learner's self concept from the Instructional Objectives Exchange in Los Angeles, California. It is designed to be administered by the teacher in the classroom setting. The teacher announces the opportunity for students to display their work [illegible] on. Sufficient room must be provided to insu[illegible] do not feel that their work cannot be display[illegible] lack of space. This measure is based on the assum[illegible] students with a positive self concept will want to display their work.

Scoring and Administration: The teacher should tell the students about posting their work in a way that seems natural to the typical classroom setting. Emphasis should be placed on the voluntary nature of the activity and the fact that work posting will not be a reward-punishment situation. Care should be taken to provide this opportunity for a variety of subject areas. The teacher totals the number of papers posted during the observational period(s) and divides that by the number of children in the class to obtain a percentage of the class that participates.

Subjects: Work Posting is suitable for children in grades K-6.

Reliability and Validity: No information available.

Comments: Since this measure is part of an objectives-items bank where there is little data feedback, little is known about how it stands up in the field. It is obvious that, if used, much more information is needed before class scores can be interpreted. It would seem that its best use would be in conjunction with a learning program or technique that is designed to change students' self concepts; however, it is vital that other measures (e.g., self report type) be used to assure the teacher or investigator that he is, in fact, measuring self concept rather than other variables which might influence a child to post his/her work.

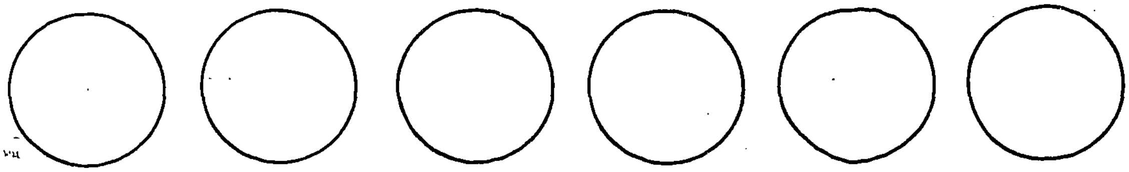
*Sample procedure reproduced by permission of W. J. Popham, Director of Instructional Objectives Exchange.

Projective Technique

Title: The Children's Self-Social Constructs Test (CSSCT)

Description: The CSSCT is a projective technique which consists of approximately 12 symbolic arrays in which circles and other figures represent the self and/or significant others and it is available in 3 forms: preschool, primary, adolescent (Henderson, Long, and Ziller, 1965). The child is required to arrange these symbols by selecting a circle to represent the self or some other person, by drawing a circle to stand for him/herself or another, by pasting a gummed picture that represents the self onto a page with other symbols, or by placing a letter in circles (e.g., M for mother) arranged on a page. The assumption underlying the instrument is that inferences can be made about a person's self concept from the ways in which the subject relates him/herself symbolically to a variety of social configurations. Each form of the CSSCT is designed to measure self esteem, social interest, identification, minority identification, realism to size, preference for others, while the primary form measures a complexity dimension as well.

Example:* Horizontal self esteem (adolescent version)



(The subject marks each circle with letter standing for a person on a list: D - doctor, F - father, Fr - friend, S - yourself, etc. Additional stimuli are presented for a new set of blank circles such as: F - someone who is flunking; K - someone who kind, S - yourself, etc.)

Scoring and Administration: Scoring is somewhat complex but the manual provides guidance for scoring each task. Each form has a different method and directions for administration (e.g., preschool form is administered individually; adolescent form in groups). All forms are administered orally. Experience and training are required to give the test.

Subjects: An early study involved 420 students in grades 6-12. Five different samples of children of school and preschool ages were tested in reliability studies. Norms for boys and girls are available. Since its development the instrument has been used in a variety of independent research endeavors.

Reliability: Four different samples ranging from grade K-12 were used to determine split half reliability coefficients (internal consist-

*Sample item reproduced by permission E. A. Henderson, B. H. Long, and R. C. Ziller, the copyright owners. Tests to be published by Educational Testing Service, Princeton, New Jersey.

ency). One sample (6th graders) was used to determine test-retest reliability. For example, for the adolescent test, split-half coefficients on 11 tasks ranged from .58 to .94. More extensive data is in the manual.

Validity: The manual carefully discusses each of the tasks in terms of theoretical grounding (content validity) and empirical findings (e.g., correlations of each of the tasks on the CSSCT with other instruments and methods for measuring self esteem). Validity coefficients must be interpreted with several factors in mind such as age of subject, ethnic background, etc.

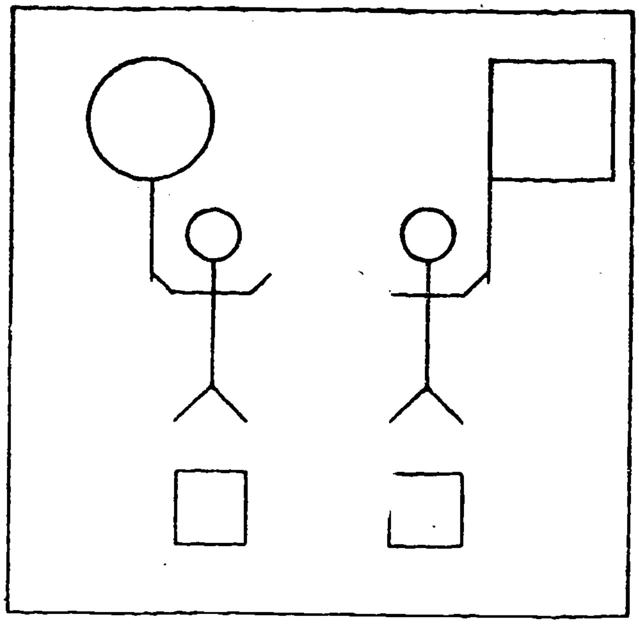
Comments: Great theoretical care has been taken in developing the CSSCT, and [illegible] research that involves self-social symbol tasks is [illegible] [illegible]sive. Projective instruments are apt to show up poorly when subject to psychometric interpretation; however, [illegible] is an exception. Since the tasks are essentially nonverbal and appear to be intrinsically interesting to children, they have wide applicability.

Semi-Projective Technique

Title: The Children's Self Concept Index (CSCI)

Description: The CSCI is a 26-item inventory designed for Project Headstart to assess the degree of positive self concept of children in grades 1-3. Peer acceptance and a positive reinforcement in the home and school are the major areas of emphasis in the index. Each item is composed of two sentences. One pertains to a balloon child, the other a flag child represented by a pair of stick figures. The child representing the socially desirable attribute is represented at alternate times by the two stick figures so that neither the balloon child nor the flag child is the good child throughout the 26 items. The problem of numbering items is eliminated by using different colored pages for each item.

Example: The administrator says, "I'm going to tell you a story. Listen carefully and mark an X in the little square under the child who is more like you." (Read item sentences.)



Most grown-ups don't care about the balloon-child.

Grown-ups like to help the flag-child.

Scoring and Administration: The test can be given without training to individuals or classroom groups. For larger groups an aid may be necessary, especially

*Sample items reproduced by permission of Westinghouse Learning Corporation.

when dealing with first graders. Directions for administration and instructions for the children are easily understood. The entire test is read to the subjects with two sample items preceding the test to help the subjects understand the format.

Subjects: The instrument was standardized on a sample of 1,900 disadvantaged children in grades 1-3 from 9 geographic areas.

Reliability: Test-retest reliability after a 2 week interval was .66, computed on a sample of 100 second grade students. The coefficient for internal consistency was .80.

Validity: Rank order correlations of scores with teacher ratings of the child's self concept ranged from .20 to .60 for 4 different classrooms.

Comments: The low test-retest reliability may be due to personality instability in the primary years. Correlations between the CSCI and other measures of self concept would add evidence toward determining validity. The use of the test with 'middle-class' samples also would be of interest. Despite these drawbacks, the CSCI represents a creative attempt to evaluate the self concept of the very young student.

More information on the CSCI may be obtained from:

Westinghouse Learning Corporation
100 Park Avenue
New York, New York

Semi-Projective Technique

Title: Responsive Self-Concept Test

Description: Designed for the evaluation of Follow Through students, this instrument measures nine psycho-social factors in children (grades 1-3): self-awareness, emotional affect, relationship with family, relationship with peers, verbal participation, approach to learning, reaction to success/failure, self satisfaction. The child receives a booklet of colored cards, each of which has a circle or square. On a larger white backing card is pasted a picture of the child taking the test. In the square is a picture of another child who is not known to the subject. If the subject is a black male, then the picture in the square must be one of a black male, etc. After a statement is read, the child is told to put an X in the circle or square on the colored card below the picture of the child to which the statement applies. A teacher's rating scale for assessing the nine factors is available for use with the instrument.

Examples:* grey sheet: Which child likes to play alone?
orange sheet: Which child does not talk very well?

Scoring and Administration: The test can be administered by the teacher to up to seven children at one time. A Polaroid camera is needed for taking full-face snapshots of the children. Directions are clear and a warm-up session is included. Information on scoring was not available.

Subjects: Information not available.

Reliability and Validity: Psychometric data on the test are not yet available.

Comments: The instrument is unique in its design and takes into account the age of the subject. Its utility will be increased once data becomes available. The theoretical basis for using the nine psycho-social factors and the pictures of like ethnic background and sex for the 'other' child is not clear. One possible problem with the scale is that it uses colored cards with the assumption that the children know colors. Therefore, it is crucial that the teacher or an assistant make certain that the children have their booklets turned to the right card.

More information on the instrument can be obtained from:

Ann Fitz Gibbon
Far West Laboratory for Educational
Research and Development
1 Garden Circle
Hotel Claremont
Berkeley, California 94705

*Sample items reproduced by permission of A. Fitz Gibbon, the author.

Behavior Trace Report

Title: Behavior Rating Form (BRF)

Description: This form was developed for use in conjunction with the Coopersmith Self Esteem Inventory. It consists of 13 simple and compound questions about behavioral self concept indicators pertaining to a child in the classroom. The teacher checks the answer on a five-point scale. Items in the BRF refer to such behaviors as the child's reaction to failure, self confidence in a new situation, sociability with peers, and the need for encouragement. The questions were developed after a series observations in and out of the classroom and repeated interviews with teachers, principals, and a clinical psychologist.

Examples:* Does the child deprecate his school work, grades, activities and work products? Does he indicate he is not doing as well as expected?

_____ always _____ usually _____ sometimes

_____ seldom _____ never

How often is the child chosen for activities by his classmates? Is his companionship sought for and valued?

_____ always _____ usually _____sometimes

_____ seldom _____ never

Scoring and Administration: The BRF is self-administered and scoring information is available from the author. The BRF provides two scores--esteem behavior and defensive behavior.

Subjects: (See Coopersmith Self Esteem Inventory)

Reliability: Cross rater reliability was determined by correlating ratings of teachers and principals (.73).

Validity: (See Coopersmith Self Esteem Inventory). The author reports that there was a general tendency for the teachers to rate girls higher; however, to correct for this systematic bias, male and female scores were scaled separately.

Comments: Here again, the BRF was used by Coopersmith as a screening device; however, it can be used effectively as a validity check on self report or projective measures (e.g., correlating scores on the BRF with the Piers-Harris Self Concept Scale). Since the use of the BRF involves a retrospective report of

*Sample items reproduced from The Antecedents of Self-Esteem, by S. Coopersmith, San Francisco, California: W. H. Freeman and Company, 1968.

behavior and not a direct observation of a child's behavior, it eliminates the problem of the child knowing that he is being observed and reacting to the observer. However, a teacher's memories of a child's actions are notoriously faulty due to the numerous opportunities for distortion and bias.

Self Report

Title: Coopersmith Self Esteem Inventory (CSEI)

Description: The CSEI is a 58-item inventory concerned with the subject's self attitude in four areas: peers, parents, school, and personal interests. The inventory was devised by Coopersmith (1967) for research carried out during 1959-1965 on the antecedents, consequences, and correlates of self esteem. Most of the items were based on items from a scale by Rogers and Dymond (1954). All the items were reworded for use with children age 8-10. Then five psychologists sorted the items into two groups--those indicative of high self esteem and those indicative of low self-esteem.

Examples:*

	Like Me	Unlike Me
I'm a failure	________	________
I'm never shy	________	________
It's pretty tough to be me	________	________

Scoring and Administration: The inventory may be group administered to persons aged 9 and older. Individual administration or rewording of the terms may be necessary with children younger than age nine. The author also has a shortened version for children in grade 3. Scoring information is available from the author.

Subjects: The inventory originally was administered to 1,748 children attending public schools in central Connecticut. It has been administered to other samples in independent studies since Coopersmith's work was published.

Reliability: Test-retest reliability after a three-year interval was .70. A five-week interval test-retest reliability study produced a coefficient of .88.

Validity: Since the CSEI was used for purposes of screening and selecting a sample for the major portion of the study, validity information is not directly available. For Coopersmith's purposes, validity is reported via the results of his study and not in terms of validity coefficients. Other evidence for validity can be found in data from other studies in which the inventory was used.

Comments: The study for which this instrument was developed is the most widely known and studied monograph on the subject of self esteem. Consequently, the instrument along with other techniques have been used by many researchers and evaluators. However, other

*Sample items reproduced from The Antecedents of Self-Esteem, by S. Coopersmith, San Francisco, California: W. H. Freeman and Company, 1968.

instruments that have been summarized here have far more psychometric data from which to judge their utility. The language and readability of the CSEI are more difficult th[illegible] which is found in other self-report measures [illegible] this collection.

Self Report

Title: Tennessee Self Concept Scale

Description: This instrument was developed by Fitts (1955) to fill a need for a scale which is simple for the subject, widely applicable, well standardized, and multidimensional in its description of the self concept. The scale consists of 100 self descriptive statements and the subject judges each statement on a five point scale. Subjects age 12 or with a sixth grade reading ability can use the TSCS. A variety of subscales are embedded in the inventory and vary as to whether the scores will be used for counseling, clinical work, or research. The TSCS is applicable to subjects in the whole range of psychological adjustment.

Examples:* I like my looks just the way they are
I find it hard to talk to strangers
I am a nobody

Completely false	Mostly false	Partly false and partly true	Mostly true	Completely true
1	2	3	4	5

Scoring and Administration: Hand scoring is a complicated procedure because of the subscales, and the author suggests the use of the available computer scoring service for 50 or more tests. The scale can be self administered for either individuals or groups.

Subjects: The standardization group from which norms were developed was a sample of 626 people. The sample included subjects from various parts of the country, from ages ranging from 12 to 68, from various ethnic groups, socioeconomic levels, and educational levels. Subsequent studies and samples showed group means and variances which are comparable to the norming sample.

Reliability: Test-retest reliability coefficients for all major subscores ranged from .61 to .92. The time interval between measurements was two weeks. Other evidence of reliability was the similarity of profile patterns found through repeated measures on the same individuals over long periods of time. The author cites that reliability coefficients for profile segments used in one of the subscores fall in the .80 - .90 range.

Validity: Validation procedures used in conjunction with the TSCS were of four kinds: (1) content validity (e.g., an item was retained in the scale only if there was unanimous agreement by a group of judges that it was classified properly in a system that was

*Sample items reproduced by permission of author, W. H. Fitts and publisher, Counselor Recordings and Tests, Nashville, Tennessee.

used to determine subscores); (2) discrimination between groups (e.g., subscores were analyzed to determine whether they differentiated between psychiatric patients and non-psychiatric patients and within patient groups in a variety of settings); (3) correlation with other personality measures (e.g., Minnesota Multyphasic Personality Inventory, Edwards Personal Preference Schedule); (4) personality changes under particular conditions (e.g., changes through psychotherapy, drug therapy and experiments). In summary, most of the procedures provided substantial evidence as to the validity of the instrument.

Comments: Recently the TSCS has been used in several studies relating self concept to school achievement. Its simple language and ease of administration are desirable in a practical setting. The extent of psychometric data in the manual and new research data add to its soundness as a measurement tool. Several drawbacks are evident. The manual and scoring procedures are somewhat complex, and the instructions to the subject are curt and test-like in tone, which hinders the establishment of comfort and rapport with the subject. It is considerably longer than other measures of self concept.

Self Report

Title: How I See Myself Scale (Primary and Secondary Form)

Description: This is a 40-item scale for the primary version (grades 3-6) and a 42-item scale for the secondary version (grades 7-12) developed by Gordon (1966) for use in a variety of research projects. The basic assumption underlying the scale is that self concept is not a unitary trait. Therefore, the scale contains several rationally derived subscales which relate to student's view of peers, teachers, school, and his/her own emotional control. Factor analytic studies produced five major factors. They were labeled Teacher-School, Physical Appearance, Interpersonal Adequacy, Autonomy, Academic Adequacy.

Examples:*
I don't like teachers 1 2 3 4 5 I like teachers very much
I'm just the right weight 1 2 3 4 5 I wish I were heavier, lighter
I don't read well 1 2 3 4 5 I read very well

Scoring and Administration: Items were randomly reversed to reduce any tendency to mark column 5 when answering the items. Scores on individual items must be converted so that 5 always represents the positive end of the scale. Scores are derived on the basis of the factors from the results of empirical studies done with the instrument. The inventory is suitable for group administration, and the directions to be read by the administrator are clear and provide for the establishment of rapport with the group. The author suggests that each item be read separately to third graders. Norms are available for grades 3-12 by sex, race, and social class.

Subjects: The inventory was developed by testing students (grades 3-12) in a laboratory school at the University of Florida. The factor analytic study resulted from collecting data from a total of 8,979 school children in a north central public school system.

Reliability: Three separate test-retest reliability studies were done on the basis of the factor scores and total scores. One included a group of "disadvantaged" mothers. Intervals between testing ranged from nine days to two weeks. Reliability coefficients using total score ranged from .87 to .89. Studies using factor scores had coefficients for factors ranging from .45 to .82.

Validity: Content validity was established by the use of a model and material from Jersild (1959) who used an open-ended composition approach and then categorized the responses of children and adolescents. The items on the inventory were based on these categories.

*Sample items reproduced by permission of I. Gordon, the copyright owner. Manual published by Florida Educational Research and Development Council, Gainesville, Florida.

Studies were undertaken to assess other aspects of validity. Inventory scores were correlated with scores from an inferential technique; an observer used a mixture of interview, projective techniques, and observation and quantified inferences on a seven point rating scale. Correlations were positive and non zero but generally low. Ratings from classroom behavior observations were correlated with inventory scores. Even though the observations covered a variety of topics and procedures, there were low but significant correlations between all parts of the scale and observed classroom behavior. Other studies included comparison of student scores with adult scores, obtained from the sample of mothers used in the reliability study.

Comments: The author admits that further work in comparing this scale with other instruments, observed behavior, and with environmental and developmental variables is necessary. However, more than the average amount of care and time have been taken in the development and study of the instrument since its inception in 1959. It is one of the few self concept inventories that comes with a manual and a rationale. It will no doubt be used in other studies.

Self Report

Title: A Semantic Differential for Measurement of Global and Specific Self Concepts

Description: This scale, a derivation of the technique described by Osgood (1957), was developed for use in research for a dissertation (Stillwell, 1965) and was used subsequently in an ESEA project to determine changes in student attitude after counseling. There are two versions of the scale--grades 1-3 and grades 4-6. On a typical semantic differential the subject rates a particular concept on several seven-step, bipolar adjective scales. For public school children, a five-step scale is recommended. The author decided to use a verbal format rather than a numerical one for the steps. Concepts used were Myself, Myself as a Student, Myself as a Reader, Myself as an Arithmetic Student. Nine bipolar adjective scales were used, differing slightly for the two forms.

Examples:*

Myself (Grades 4-6)

very useful	somewhat useful	average	somewhat useless	very useless
very strong	somewhat strong	average	somewhat weak	very weak

Myself (Grades 1-3)

weak	average	strong
sad	average	glad

Scoring and Administration: Scoring is accomplished by assigning numbers 1 through 5 for each adjective pair, resulting in a possible total of 45 for each concept. This is, of course, different for the primary form, which has a possible total of 27. The scale is easily administered to entire classes, and warm up time is given in the form of rating sample concepts which are unrelated to self-esteem. There are administration problems with very young children; therefore, the author suggests that with first and second graders several assistants should be used to help children keep their places and "read" the items.

Subjects: In the original study, 230 sixth grade students completed the forms. Means and standard deviations are available for this group. However, there are no comprehensive normative data.

*Sample items reproduced by permission of author, L. Stillwell Corbett.

Reliability: Reliability coefficients are reported in terms of test-retest data (.47 - .92 for girls and .57 - .71 for boys) and on the method of rational equivalence which is a measure of internal consistency (.55 - .90 for girls and .63 - .85 for boys).

Validity: Scores on the Myself and Myself as Student scales were correlated with eight questions pertaining to self concept as a student from the Coopersmith (1959) self-esteem inventory and with scores on a behavior rating form (also by Coopersmith) filled out by the students' teachers. Substantial coefficients were obtained. It was not possible to find other methods or instruments relating to Myself as a Reader and Myself as an Arithmetic Student. However, when scores on these were correlated with scores on Myself and Myself as Student, the intercorrelations showed that each scale measured a different aspect of self concept.

Comments: Although this particular instrument has not been used widely, measuring self concept with the semantic differential technique has been done in a variety of settings. It is an economical and practical method of gathering data. Verbal content is at a minimum, and, therefore, the instrument eliminates the problem of gathering information from the young child or the poor reader.

Self Report

Title: The Piers-Harris Children's Self Concept Scale (The Way I Feel About Myself)

Description: This inventory is an 80-item instrument designed primarily for research on the development of children's self attitudes and correlates of these attitudes (Piers and Harris, 1964). It was thought that when deriving items for the scale, the universe to be sampled in a children's self concept measure should consist of items reflecting the concerns that children have about themselves; therefore, the authors used Jersild's (1952) collection of children's statements about what they liked and disliked about themselves. The items are simple declarative statements with at least half being negative in content. Subjects are to circle "yes" if the item is true for them and "no" if it is not true. The test is suitable for children in grades 3-12.

Examples:*

I am dumb about most things	yes	no
I am good in my school work	yes	no
My parents expect too much of me	yes	no

Scoring and Administration: Scoring is simple with 1 = yes and 0 = no for a maximum score of 80 on the inventory. The author recommends that the inventory be administered orally to grade 6 and below. Children below age eight or third graders should receive individual administration. No training is necessary to give the test, and instructions provide for the establishment of rapport with the subjects.

Subjects: The instrument was normed on a sample of 1,183 public school children in a Pennsylvania school district in grades 4-12. From 1964 to 1967 it was used in nine studies involving children from different parts of the U.S. and from different groups such as special education students, stutterers, economically deprived, etc.

Reliability: Internal consistency coefficients ranged from .78 - .93 using the KR-21 formula; however, when the Spearman-Brown formula was applied, the range was .87 - .90. Test-retest coefficients after a four month interval ranged from .71 - .77.

Validity: At the outset of the instrument's development, content validity was considered by using Jersild's (1952) data. Scores on the Piers-Harris scale have been compared with other self concept measures resulting in reasonably high validity coefficients. Teacher and peer ratings correlated with the scale produced coefficients ranging from .06 to .49. Ratings of other variables such as socially effective behavior and superego strength were also compared to the scores on the Piers-Harris.

*Sample items reproduced by permission of the authors, D. B. Harris and E. V. Piers.

Factor analysis of the scale revealed six major factors which were labeled Behavior, Intellectual and School Status, Physical Attributes, Anxiety, Popularity, Happiness, and Satisfaction.

Comments: The Piers-Harris Scale is commercially produced and has been used widely in educational evaluation and research. It is superior to most self-report, paper-and-pencil procedures for self concept in that psychometric data is available, and its use in ongoing research adds evidence as to its validity. It is accompanied by an excellent semi-technical manual. More information can be obtained from:

Counselor Recordings and Tests
Box 6184 Acklen Station
Nashville, Tennessee 37212

Self Report

Titles: Michigan State General Self Concept of Ability
Michigan State Self Concept of Ability in Specific Subjects Scales

Description: The Michigan State University instruments were devised by Brookover, Patterson, and Thomas (1962) for a USOE Cooperative Research Project and were used in a subsequent experimental research project in Michigan in 1965. The general version attempts to measure the evaluation one makes of oneself with respect to the ability to achieve in academic tasks in general as compared to others. This inventory consists of eight items each coded from 5 to 1. The specific form measures the evaluation one makes of oneself in respect to a given subject matter area. The items for these scales are directly parallel to items in the general instrument. Both measures are suitable for students in grades 7-12.

Examples:

General

How do you rate yourself in school ability compared with your close friends?

a. I am the best
b. I am above average
c. I am average
d. I am below average
e. I am the poorest

Specific

How do you rate your ability in the following school subjects compared with your close friends?

	among the poorest	below average	average	above average	among the best
Mathematics	□	□	□	□	□
English	□	□	□	□	□
Social Studies	□	□	□	□	□
Science	□	□	□	□	□

Scoring and Administration: In the general form, the higher the self concept the higher the numerical value on each item with 40 being the maximum score. Scoring is essentially the same in the specific form except that

*Sample items reprinted by permission of W. B. Brookover.

each question involves four subject areas thus giving four, eight-item tests which are scored like the general form. The instruments are self-administered and designed for group administrations.

Subjects: Approximately 1,500 white students in an urban school setting in grades 4-10 were tested in the course of the two USOE Cooperative Research Projects. The instruments have been used in other research, sometimes in a revised form.

Reliability: The eight item general form produced test-retest coefficients of .75 for males (n = 446) and .77 for females (n = 508) after a year's interval. Internal consistency coefficients ranged from .82 - .92 for males and .77 - .84 for females with large samples of students in grades 7 - 10. The general form has the characteristics of a Guttman scale with high coefficients of reproducibility. The specific form showed test-retest correlations from .63 - .80 and internal consistency coefficients in ranges similar to the general form.

Validity: The general self concept of ability scale was correlated with a variety of variables (e.g., evaluations of teachers, friends, parents; grade point average; scores on specific self-concept of ability). This instrument showed consistently high correlations with the other variables.

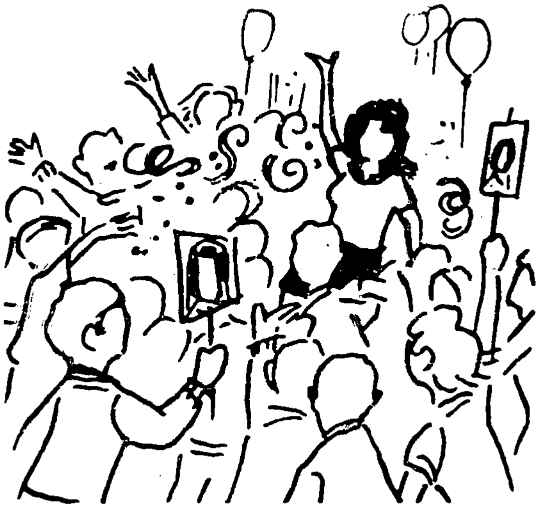
Comments: These instruments are unusual in that they focus on one differentiated aspect of the self concept--academic ability--whereas most other self concept measures consider several aspects of self concept. Studies relating other aspects of self concept and self concept of ability would add to validity information. An interesting side benefit from the study was the discovery that the older student's evaluation of him/herself as a student is a realistic one and not subject to faking. Recent studies by other researchers have shown that a student's evaluation of him/herself and his/her self reports of grades predict success in college (freshman grade-point average) as well as placement tests and actual high school grade-point average.

Pictorial Self Report

Title: Self Esteem Measure for Neighborhood Youth Corps Enrollees

Description: This 16-item inventory consists of pictorial scenes in which the adolescent is portrayed in various academic, social, and employment settings and is one of a varied battery of measures which assess work behavior. The subject is asked to imagine that the young person in the picture represents him/herself. The subject's response on a three point scale is intended to reflect his/her level of self-worth. The measure was developed by Freeberg (1968) for a Department of Labor project after he rejected a group of published measures because they appeared to be unsuitable for a disadvantaged adolescent group.

Example:*



□ I'm the kind of girl who can be leader and who people look up to - like in this picture.

□ I could never be like that girl in the picture with people cheering me.

□ I might be good at some things that people would look up to me for.

Scoring and Administration: The total score on the scale is obtained by summing all item weights where the weights are 1-3 on each item with 3 representing the high point of the continuum. The measure is intended for administration to small groups with a maximum of 10 individuals per group. There are separate tests for males and females. Directions and all item stems and choices are read to the subjects.

Subjects: The scale was administered to 133 males and 133 females from rural and urban areas who were Neighborhood Youth Corps enrollees in [illegible] centers in the northeast and southeast United States.

Reliability: Internal consistency coefficients served as estimates of reliability. They were .50 for males and .60 for females.

Validity: A validity study correlated scores on the measure with counselor and work supervisor's criterion ratings. Coefficients for male

*Developed by Educational Testing Service for the Neighborhood Youth Corps (NYC) under a contract with the U.S. Department of Labor.

enrollees were very low (.04 and .01) and slightly higher for females (.15 and .21). Factor analysis of the entire battery of scales showed that one of the features of the self-esteem scale is the relatively "pure" attitudinal aspect of its contribution to the battery.

Comments: Reliability estimates may be low because of the brevity of the scale. Unfortunately, this may have contributed heavily to lowering the validity coefficients. However, the measurement technique could be quite useful. A pictorial instrument which is relevant to adolescent experience is missing from any of the lists of school-oriented self concept measures.

REFERENCES

Brookover, W. B., Patterson, A., & Thomas, S. Self concept of ability and school achievement. Final Report of Cooperative Research Project No. 845, U. S. Office of Education. East Lansing, Mich.: Office of Research and Publication, Michigan State Univ., 1962.

Campbell, D. T., & Fiske, D. W. Convergent and discriminant validation by the multitrait-multimethod matrix. Psychological Bulletin, 1959, 56, 81-105.

Coller, A. R. The assessment of self concept in early childhood education. ERIC Clearinghouse on Early Childhood Education, July 1971.

Coopersmith, S. A method for determining types of self esteem. Journal of Abnormal and Social Psychology, 1959, 59, 87-94.

Coopersmith, S. The antecedents of self esteem. San Francisco, Calif.: W. H. Freeman and Co., 1967.

Creelman, M. B. The CSC test manual, 1954 (available from author, Department of Psychology, Cleveland State University, Cleveland, Ohio).

Fitts, W. H. Preliminary manual, the Tennessee Department of Mental Health self concept scale. Nashville, Tenn.: Department of Mental Health, 1955. (Mimeographed)

Freeberg, N. E. Development of evaluation measures for use with Neighborhood Youth Corps enrollees. Final Report, U. S. Department of Labor, Contract #66-00-09. Princeton, N.J.: Educational Testing Service.

Gordon, I. Studying the child in school. New York: John Wiley & Sons, 1966.

Gordon, I. A test manual for the How I See Myself Scale. Gainesville, Fla.: Florida Educational Research and Development Council, 1968.

Henderson, F. H., Long, B. H., & Ziller, R. C. Self-social constructs of achieving and nonachieving readers. The Reading Teacher, 1965, 19, 114-118.

Jersild, A. T. In search of self. New York: Teachers College Bureau of Publications, 1952.

Osgood, C. E., et al. The measurement of meaning. Urbana, Ill.: Univ. of Illinois Press, 1957.

Piers, E. V., & Harris, D. B. Age and other correlates of self-concept in children. Journal of Educational Psychology, 1964, 55, 91-95.

Rogers, C. R., & Dymond, R. F. (Eds.) Psychotherapy and personality change: coordinated studies in the client-centered approach. Chicago, Ill.: University of Chicago Press, 1954.

Stillwell, L. An investigation of the interrelationships among global self concept and achievement. Unpublished doctoral dissertation, Western Reserve Univ., 1965.

Wylie, R. *The self concept*. Lincoln, Neb.: Univ. of Nebraska Press, 1961.

Zirkel, P. A. Self concept and the 'disadvantage' of ethnic group membership and mixture. *Review of Educational Research*, 1971, 41, 211-215.

AN OMNIBUS OF MEASURES RELATED TO SCHOOL-BASED ATTITUDES

Joan Knapp

TABLE OF CONTENTS

Introduction

The following summaries form a small sample drawn from a burgeoning corpus of literature concerning the measurement of school-based attitudes. These measures were selected to show the variety of instruments available. Some involve pictorial stimuli; others involve written statements. Some are projective in nature; others are more objective. Some are designed for children, ages 4-8; others for students in grades 12-14. Some have evolved through several decades of research; others have virtually no published data which would contribute to the evaluation of their soundness.

The amount and variety of these measures indicate that there is an increased interest in assessing students' attitudes. However, an examination of the following sample from the literature reveals that there are many closely related terms in this non-cognitive realm which prevent the educational researcher and evaluator from attaining a tight conceptualization of this area. Attitude can mean opinion, feelings, habit, self-concept. School can mean learning, study, education, teachers, or a particular subject such as mathematics.

For this reason, it is imperative that the education specialist clearly defines his goals and objectives in the affective-attitudinal domain before borrowing an instrument designed by others or before designing one himself. Such goals and objectives should be embedded in a larger conceptual scheme which includes other variables that are salient to his purpose and that add to the validity of his instrument. For example, an evaluation specialist may be interested in changes in attitudes toward mathematics on the junior high level. Depending on his purpose, interpretation of results on a student mathematics attitude questionnaire may involve measuring home environment variables, attitudes and training of teachers, the content and goals of various math curricula being used, etc. In other words, the self-report inventory cannot be the "be all" and "end all" when attempting attitude assessment.

As varied as the following collection seems, all the instruments have one characteristic in common. They are all paper-and-pencil, self-report inventories and suffer from all the inherent disadvantages of this measurement technique. They are subject to malingering and faking on the part of the student. Response sets and styles may introduce much error in the measurement procedure. These possibilities serve to threaten the validity of attitude inventories. Attitudinal behavior as measured by questionnaires is more changeable over time than cognitive behavior as measured by tests of ability. This complicates the determination of reliability. Since most of the inventories involve reading and some involve writing, the student's ability to answer the items accurately is largely dependent on his or her verbal aptitude. This problem is particularly critical at the elementary school level. For these reasons, self-report inventories should be supported by attitudinal data obtained from other measurement techniques such as observations, interviews, peer and teacher ratings, school records, and so on.

As a final comment, much of the research in this area relies heavily on correlational techniques applied to these indirect somewhat crude measures of affective behavior. Not enough expertise has been directed to testing out hypotheses suggested by correlational research. If the educator's goal is to change cognitive and affective behavior, then more sophisticated research and assessment is needed to discover techniques for developing positive attitudes and modifying negative attitudes.

Title: Survey of Study Habits and Attitudes

Description: The SSHA is a 100 item instrument developed by Brown Holtzman (1955) with three purposes in mind. They are 1) to identify students whose study habits and attitudes are different from those of students who earn high grades, 2) to aid in understanding students with academic difficulties, and 3) to provide a basis for helping such students improve their study habits. The current Form C (Grades 12-14) and Form H (Grades 7-12) are based upon eight years of research and development. The authors claim that the instrument has four subscales--delay avoidance, work methods, teacher approval, educational acceptance.

Examples:

1. I lose interest in my studies after the first few days of a semester.
2. When I am having difficulty with my school work, I try to talk over the trouble with the teacher.

Subjects: The instrument, both forms, has been validated and tested on thousands of college and secondary school students. For example, Form H was normed on a total of 11,218 students in 16 different towns and metropolitan areas across the United States.

Response Mode: For each item the student blackens one space out of five, marked R, S, F, G, A, which correspond to rarely, sometimes, frequently, generally, almost always.

Scoring: The alternatives are scored from 5 (almost always) to 1 (rarely) for positively phrased items. Weights for negatively phrased items are reversed. The student's score is the sum of the weights for alternatives endorsed by him. High scores indicate more positive attitudes and habits. In addition, subscores can be obtained for counseling purposes.

Comments: The SSHA has been used in many research studies and it has been reviewed in Buros' Mental Measurement Yearbook. The subscales were derived empirically; whereas most instruments of this type have somewhat weaker subscales derived rationally. SSHA is published complete with manual by the Psychological Corporation.

Title: School Interest Inventory

Description: This instrument was designed to identify potential dropouts (Cottle, 1961). It consists of 150 statements. Items which consistently differentiated between dropouts and stay-ins were selected from an item pool to form the present inventory.

Examples:

1. I like school.
2. I skip school at least once a month.
3. I have been absent from school more than 20 days this year.

Subjects: The instrument has been administered to 25,000 students in ten states after being validated on a matched sample of 1,300 dropouts and 1,300 stay-ins. For maximum usefulness the inventory should be administered to junior high school students.

Response Mode: Subjects are asked to fill in circle containing "T" if the item is true for him; fill in circle containing "F" if item is false.

Scoring: Unweighted and weighted scores can be determined, with an absolute unweighted raw score of 30 suggested as a cutting score above which students might be considered dropouts. Of the items, 75 are scored for both males and females, 15 for just males, 11 for just females, and 49 of the items have no function in determining a subject's score.

Comments: The items are transparent, thus promoting faking. Those items which receive the greatest weights are the most obvious. The predictive strength of this measure has not been compared to the strength of attendance records, grades, or teacher opinion. The inventory is published by Houghton Mifflin Co.

Title: The Student Opinion Poll II

Description: This is a revision of a questionnaire developed by Jackson & Getzels (1959) and used by Jackson and Lahaderne (1967) in a study of 300 sixth grade students in a working class suburb. Its intent was to elicit responses concerning general satisfaction or dissatisfaction with four aspects of school life: the teachers, the curriculum, the student body, and classroom procedures. This version contains 47 multiple choice items.

Example: The things that I am asked to study are of

a. great interest to me.
b. average interest to me.
c. little interest to me.
d. no interest to me.

Subjects: Various versions of this inventory have been used in research studies involving private and public, urban and suburban, and junior and senior high school students. Adaptions would have to be made for students with poor reading skills.

Response Mode: A student indicates his response by circling the choice which best completes the item stem.

Scoring: The questionnaire is scored by giving one point each time the subject chose from within a set of multiple choices the response indicating the highest degree of satisfaction with that aspect of school life. Thus, the possible range of scores was from 0 to 47.

Comments: When used in research studies, student scores on the instrument showed no relationship to the scholastic performance of the students; however, this does not mean that the instrument cannot be used to assess the effect on school attitude of an innovative educational program.

Title: School Morale Scale

Description: The School Morale Scale (Wrightsman, Nelson, and Taranto, 1968) is an 84-item scale which measures seven aspects of a student's morale about school. These aspects ranged from morale about school plant to general feeling about attending school. Several persons independently composed statements for subscales. A total of 150 items were obtained and were reduced to 12 items for each of the seven subscales.

Examples:
1. This building is old and run-down.
2. All my teachers know me by name.
3. The principal of this school is very fair.

Subjects: The sample upon which the scale was constructed was 127 fifth graders from public elementary schools in a small city in Tennessee, 169 seventh graders from a junior high school in a large city in Tennessee, and 137 ninth graders from the same junior high school. It has been administered to fifth and sixth graders in Alabama.

Response Mode: Subjects respond by marking items with which they agree with an "A" and items with which they disagree with a "D".

Scoring: Each subscale is scored with a total of 12 indicating good morale in regard to that aspect. The scores for the seven subscales are summed to give a total score which ranges from 0 to 84.

Comments: Reliability and validity information can be obtained from authors at George Peabody College for Teachers. It is not clear that subscales are justified. The reading level of the items may prove difficult for elementary students.

Title: Measures of School and Learning Attitudes

Description: These two measures (25 items each), Attitude Toward School and Attitude Toward Learning were developed by Roshal, Frieze, and Wood (1971) for a study in which they hoped to validate these measures using the Campbell and Fiske multitrait-multimethod method. They hypothesized that attitudes toward school and learning were two separate but similar dimensions with attitude toward school being feelings about school, teachers, subjects, classmates, etc. and attitudes toward learning being concerned with the student's general interest in the world, reading, and learning activities. A third scale, Attitude Toward Technology, was devised to prove that the other two dimensions were quite different from the third.

Large numbers of items were constructed on the basis of content validity (items believed by educational specialists to measure the respective attitude) for each of the three scales. After several preliminary item analysis studies, which utilized factor analyses and item-total correlations, the final versions were constructed. Both positively and negatively worded items were used to control for response bias.

Examples: ATTITUDE TOWARD SCHOOL

1. I
a. always
b. usually
c. sometimes
d. rarely
e. never
hate school

2. Teachers in this school are
a. always
b. usually
c. sometimes
d. rarely
e. never
friendly.

ATTITUDE TOWARD LEARNING

1. School subjects are
a. always
b. usually
c. sometimes
d. rarely
e. never
boring.

2. Whenever I go on a trip, I learn
a. lots of
b. many
c. some
d. a few
e. no
new things.

ATTITUDE TOWARD TECHNOLOGY

1\.
a. strongly agree
b. agree
I. c. partly agree, partly disagree that most new inventions help people live better.
d. disagree
e. strongly disagree

2\.
a. always
b. usually
I could c. sometimes learn how to fix almost anything.
d. rarely
e. never

Subjects: The three scales, ATS, ATL, and ATT, were administered along with other questionnaires and peer ratings to a sample of 610 sixth grade students in 13 public schools. Their average Lorge Thorndike verbal IQ was 101.4 with a standard deviation of 15.7. There were approximately equal numbers of boys and girls. The sample ranged from lower middle to lower upper class in socio-economic status (as judged by school district personnel).

Response Mode: The student circles the option that best completes the statement according to his own feelings.

Scoring: Information not available.

Comments: ATS, which measures the student's general attitude towards school as an institution, might be used by educators to measure feelings about school. It is probably relatively sensitive to attitude changes (although further studies of this are needed). The results of the Roshal study does give support for the independence of the two instruments even though both teachers and researchers have some difficulty differentiating the two concepts. ATL, which indicates a more general orientation toward learning, probably does reflect more of a personality trait than does ATS and thus may not be as susceptible to short term changes as are attitudes toward school.

Both scales may be administered independently or in combination for elementary school assessment. They are presently being used with children in third through fifth grades as well as with children in the sixth grade. Although the reading difficulty of words used on the scales was purposefully kept low, use with average readers below the fifth grade is not recommended, however, unless the items are read aloud. Normative data for sixth grade pupils are available from the authors.

Title: Attitudes Toward Education

Description: This is a 34 item, Thurstone type scale developed by Glassey (1945) to measure attitudes toward the value of education and the effects of education upon people. There are enough items in the scale to create several shorter equivalent forms.

Examples:

1. We cannot become good citizens unless we are educated.
2. Too much money is spent on education.
3. Education does more harm than good.

Subjects: Approximately 300 British grammar school children, ages 11-18 and their parents were used to construct the scale. Forty judges were used to sort the items for determining scale values.

Response Mode: A student checks those items with which he fully agrees and places a cross in front of those items with shich he does not fully agree. He may place a question mark in front of the item if he is totally unable to decide.

Scoring: The student's score is the median of the scale values of the items marked as 'fully agreed'. Low scores indicate positive attitudes toward education.

Comments: Although the scale was developed with British students, the language of the items seems satisfactory for use with American samples. An advantage of the scale is that it may be used with a wide range of ages and educational levels. It would appear to be most useful in identifying potential dropouts because it seems to reveal feelings of alienation from the educative process.

Title: Politte Sentence Completion Test

Description: The Politte Sentence Completion Test (PSCT) is a projective psychological test instrument for eliciting personality data from the examinee. Thirty-five sentence stems are used. It can serve as an addition to other diagnostic and evaluative instruments used in personality assessment in the educational, counseling, and clinical areas. Ten of the items refer to attitudes toward school and school adjustment. It can be used in a 1:1 setting or in a group setting. Little training is required to administer the test; however, only qualified school or clinical psychologists should attempt to interpret the test because of the projective quality involved.

Examples:

1. What bothers me at school is ____________________.
2. School would be better if ____________________.

Subjects: The PSCT is designed for use with students in grades 7 through 12 and can be used with older subjects who are functioning at this school level. The instrument was not designed with the use of a sample; therefore, typical instrument construction data are not available.

Response Mode: Students are to complete each stem according to the way they feel about the item.

Scoring: The PSCT is not scored objectively. It can be analyzed subjectively through the use of psychodiagnostic theories involving projective techniques. Persons without training in clinical psychology should use the PSCT as a screening instrument to aid in the interviewing or counseling. Clinically trained psychologists can additionally base their interpretations on a psychoanalytic, social, behavioral or similar approach.

Comments: The subject's responses to the items are dependent on his written verbal aptitude. Because of the projective nature of the instrument, it is probably appropriate for individual counseling and would not be useful in a large group assessment situation. The test is published by Psychologists and Educators Inc., Jacksonville, Illinois.

Title: Children's Attitudinal Range Indicator

Description: This instrument is one of a battery of measures designed to assist in the study of personality factors and their relationship to achievement. They were developed particularly for preschool and early elementary students who might be characterized as culturally different (Cicirelli *et al*, 1971).

The Children's Attitudinal Range Indicator (CARI) was designed to assess the child's positive and negative attitudes toward peers, home, school, and society. In attempting to assess attitudes of the primary school child, the usual methods of attitude measurement are not applicable because young children often cannot or will not verbalize. With this in mind, a semiprojective device was developed.

The projective feature of the CARI consists of presenting unstructured and incomplete picture stories in three "frames," with a fourth frame containing three stylized conventional faces depicting happy, neutral, or sad feelings. By having the subject indicate how each story should end, the CARI invites his identification with the character of a particular frame series, his investment of self in the situation presented, and a projection of his own thinking, feeling, and judgment to determine the outcome. Thus, for example, a given item presents three frames showing Bobby on his way to school, approaching the building, and going inside; the subject is then required to choose which of the three faces is Bobby's. When he is asked to identify himself with Bobby, the child presumably projects himself into this situation and chooses the response for Bobby that reveals his own attitude towards school.

The CARI consists of eight picture stories in each of four areas (school, home, peers, and society), making a total of 32 items.

Examples:

1. (Peers) Sally is at school. A new girl comes to the class. At recess, the new girl comes over to talk to Sally. Which one is Sally's face?
2. (School) Eobby is on his way to school. He gets to school. He opens the door and goes inside. Which one is Bobby's face?

Subjects: Approximately 150 lower- and middle-class second grade pupils were used to determine the validity and reliability of the instrument. The pictorial content of the test makes it suitable for students in grades K-3.

Response Mode: Students are to circle the face that indicates how the story should end.

Scoring: Response alternatives to each of the items in the CARI are scored from one to three points; three reflects a more positive attitude, two a neutral attitude, and one a negative attitude. Subscores ranging from eight to twenty-four for attitudes towards peers, school, home, and society are obtained by adding the scores on each of the eight items representing a particular area.

Comments: Scores on the CARI were obtained in connection with a nationwide evaluation of Head Start centers; however, the instrument seems suitable for any preschool or early elementary school pupil. It should be noted that the semiprojective nature of the items encourages spontaneous responses; such responses may lower the reliability and validity of the instrument.

Title: When Do I Smile?

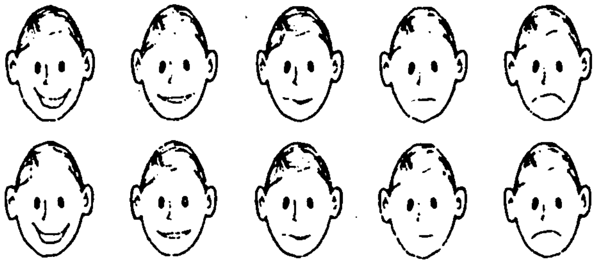
Description: This 23-item inventory was developed by American Institutes for Research to be used as one method of evaluating the attitudinal variables involved in a special innovative program in a school system in Florida. It was hoped that the pupils in the program would develop more positive and realistic attitudes toward themselves and the world. Approximately 14 items out of the total inventory concern the students' feelings toward school. Each item is accompanied by five faces depicting a range from "very happy" to "very unhappy." Separate forms were developed for grades 1-3 and grades 3-5.

Example: From Grades 3-5 form:

VERY HAPPY   HAPPY   IN-BETWEEN   UNHAPPY   VERY UNHAPPY

1 HOW DID YOU FEEL ABOUT COMING TO SCHOOL THIS MORNING?

2 HOW DO YOU FEEL ABOUT THE BOYS AND GIRLS IN THIS CLASS?



Subjects: The scale was administered to 1,616 students in grades 1-5 who had participated in the special innovative program. Developed specifically to assist in this program's evaluation, it has not been used in other situations or locations.

Response Mode: The Grades 1-3 form is designed to be administered orally. Students mark an X on the face that corresponds to the way they feel about each question. The wording of the two forms is similar. However, students using the grades 3-5 form are required to read the items themselves.

Scoring: The faces for each item represent a score range of 1-5 with 5 being "very happy." Total score is obtained by adding the score for each item. For program evaluation, the inventory was administered at the onset of the program, at the end of the program, and at the end of the school year to obtain difference scores to ascertain whether there was improvement, impairment, or no change in attitudes.

Comments: Researchers involved in the instrument's development feel that much more research and development is needed if it is to be used in other evaluations. They feel that any self report inventory for children of these ages is quite sensitive to differences in administration, perceived social and economic status, and so forth. In addition, the instrument may appear juvenile to the mature fifth graders.

Title: Attitude Toward Any School Subject

Description: This 45-item, Thurstone-type scale developed by Silance and Remmers (1934), includes two equivalent forms, which adds to its usefulness for research purposes. The inventory can be used by substituting the name of the subject under study for the words this subject in each item.

Examples:

1. This subject fascinates me.
2. My parents never had this subject, so I see no merit in it.
3. This subject does not teach you to think.

Subjects: The exact populations and samples upon which the scale was constructed are uncertain, but the sample apparently involved several thousand high school students and college undergraduates.

Response Mode: Students check those items with which they agree.

Scoring: The individual student's score is the median of the scale values (previously determined by construction sample) of the items endorsed by the student.

Comments: Even though it was developed nearly 40 years ago, this scale is still widely used in a variety of research projects. Measurement specialists feel that it is reasonably valid and reliable; however, the reading level is probably too high for a poor junior or senior high reader and the terms used in some items are somewhat dated (for example, fogy, bunk, hate it like the plague).

Title: Attitude Instrument to Evaluate Student Attitudes toward Science and Scientists

Description: This instrument was designed by Mots (1972) to determine the attitudes of sixth and ninth grade rural, urban, and suburban students toward science and scientists. The attitude instrument was based on a grid of key statements about science and scientists. Part I of the instrument consisted of statements about science, Part II about scientists.

Ideas and statements about science and scientists were obtained by questioning 525 elementary, secondary, and college students as well as scientists and science educators. The final form of the instrument was developed after extensive trial administrations for readability and understanding of the attitude statements. The instrument was validated by a jury panel of twenty professional scientists and science educators.

Subjects: The instrument was administered to a sample consisting of 981 sixth and ninth grade students from rural, urban, and suburban communities in Michigan.

Response Mode: Information not available.

Scoring: Information not available.

Comments: The instrument was designed for a particular school system in Michigan. Because of the care taken in its construction, however, it may be useful in other locations. Validity studies and replication would add to its usefulness.

Title: Inventory of Reading Attitude

Description: This 25-item instrument by Dubois (1971) attempts to assess a student's attitude toward reading in school as well as reading in free time away from school.

Examples:

1. Do you think that most things are more fun than reading?
2. Do you like to read aloud for other children at your school?

Subjects: The sample upon which the scale was originally constructed is unknown, but it has been used with elementary school children to assess the development of favorable attitudes toward reading that result from particular methods of instruction.

Response Mode: The students read each item, or the items are read to the student. The student checks "yes" or "no" for each question.

Scoring: Information not available.

Comments: The items are written simply and geared for young children. They are probably too transparent for the older child who identifies "liking reading" as socially desirable.

Title: A Childhood Attitude Inventory for Problem Solving

Description: The Childhood Attitude Inventory for Problem Solving (CAPS) was developed by Covington (1966) as part of a larger effort to develop an omnibus set of instruments to asеess problem-solving competency among upper elementary school children. CAPS is a group-administered paper-pencil inventory consisting of two 30-item scales. Scale I, which assesses the student's beliefs about the nature of the problem-solving process, treats a number of themes including the child's conception of the innateness of problem-solving ability. Scale II, which assesses the child's degree of self-confidence in dealing with problem-solving tasks, reflects some of the typical sources of childhood anxiety about thinking including the fear of having one's ideas held up for ridicule.

Subjects: A preliminary form was administered to 190 fifth and sixth grade students. The present form was administered to 325 additional subjects.

Response Mode: Information not available.

Scoring: Information not available.

Comments: The author claims that CAPS holds promise as a tool for investigating the relationship between problem-solving attitudes and various kinds of learner characteristics. Other research exploring the relationship between expressed attitudes toward problem solving and actual problem-solving performance is now being conducted.

Title: Mathematics Attitude Scale

Description: This instrument is a 20-item scale developed by Aiken (1972) using Likert's method of summated ratings. The items were derived from paragraphs written by 310 college students. Ten of the items connote positive attitudes and ten negative attitudes.

Examples:

1. Mathematics is fascinating and fun.
2. It makes me nervous to even think about having to do a math problem.

Subjects: Various versions of this scale have been used with sixth graders, junior and senior high school students, and college undergraduates and graduate students. Validity estimates were based on a sample of 160 female college sophomores.

Response Mode: Using five alternatives ranging from "strongly disagree" to "strongly agree," the student is to indicate the extent of agreement with the attitude expressed in each statement. The alternative "undecided" is included.

Scoring: The alternatives for positive items are weighted 4 (strongly agree) to 0 (strongly disagree). These weights must be reversed for negative items. The student's score is the sum of weighted alternatives endorsed by him. High scores reflect positive attitudes.

Comments: The instrument has been used by Aiken and others in extensive investigations concerning attitudes and achievement in mathematics. Variables such as age, sex, and SES have been included in the studies. The validity and reliability of this scale vary somewhat with grade level. It is generally more sound psychometrically in high school and college probably because 1) attitudes become more stable with maturity, and 2) the degree of self-insight and conscientiousness with which students can express their attitudes increases with age.

Title: A Semantic Differential for Measuring Attitudes of Elementary School Children toward Mathematics

Description: This particular instrument was developed by Scharf (1971); however, the Semantic Differential can be adapted to a wide variety of attitudinal studies. The subject is asked to indicate his response to a given concept by using a series of bipolar word pairs or antonyms such as Good versus Bad. Working fairly rapidly to heighten affective response and minimize cognitive response, the subject checks one of the positions on the scale between the pair of bipolar adjectives. The checking operation provides a series of ratings of a given concept. The same set of scales is used in rating all the concepts in the instrument.

Example: TAKING A MATH TEST IS

very : sort of : neither : sort of : very

BAD _______ : _______ : _______ : _______ : _______ GOOD

HAPPY _______ : _______ : _______ : _______ : _______ SAD

A student with a negative attitude toward "Taking a Math Test" might rate it as: very BAD and sort of SAD.

The number of concepts to be included in a particular instrument is limited only by factors of relevance and time constraints. The following concepts were included in the instrument.

1. My Math Class is
2. Doing Math is
3. Taking a Math Test is

The student's attitude toward the study of mathematics can thus be broken down into a number of component parts related to various experiences in mathematics.

Subjects: The instrument was administered in 1969 to fourth, fifth, and sixth grade students in four schools in which students had been exposed to an individually prescribed math instruction program for three years and in four control schools where traditional math was taught. A total of 1,304 students participated.

Response Mode: See above

Scoring: The directions of the student's attitude toward a particular concept, favorable or unfavorable, is indicated by his judgments within the polar terms. The intensity of the attitude is indicated by how far the score lies from the midpoint; that is, a student could respond that "Taking a Math Test is" Very Good or Sort of Good and the first response would indicate a more intense and positive attitude toward the concept than would the second. A total score can be obtained by adding up scores on the particular concepts. However, there is some question as to how meaningful this total score is.

Comments: The scales are relatively easy to construct and analyze. Such an instrument represents an attempt to construct items for attitude measurement that are comprehensible and yet are not transparent to young children.

REFERENCES

Aiken, L. R. Research on attitudes toward mathematics. Arithmetic Teacher, 1972, 14, 229-234.

Brown, W. F. & Holtzman, W. H. A study attitudes questionnaire for predicting academic success. Journal of Educational Psychology, 1955, 46, 75-84.

Cicirelli, V. G., et al. Measures of self-concept, attitudes, and achievement motivation of primary grade children. Journal of School Psychology, 1971, 9, 383-392.

Cottle, W. C. The school interest inventory. Psychological Reports, 1961, 9, 66.

Covington, M. V. New directions in the appraisal of creative thinking potential. Univer. of California, Berkeley, 1966. (mimeographed)

Dubois, J. A comparison of reading attitude between first grade reading instruction in i/t/a and t/o. Contemporary Education, 1971, 42, 169-172.

Farr, R., Laffey, J. & Smith, C. Inventory of reading attitude. In Taxonomy of evaluation techniques for reading programs. Indiana Univer., August, 1968.

Glassey, W. The attitude of grammar school pupils and their parents to education, religion, and sport. British Journal of Educational Psychology, 1945, 15, 101-104.

Jackson, P. W., & Getzels, J. W. Psychological health and classroom functioning: a study of dissatisfaction with school among adolescents. Journal of Educational Psychology, 1959, 50, 295-300.

Jackson, P. W. & Lahaderne, H. M. Scholastic success and attitude toward school in a population of sixth graders. Journal of Educational Psychology, 1967, 58, 15-18.

Mots, L. L. The development of an instrument to evaluate sixth and ninth grade students' attitudes toward science and scientists. Proceedings of 45th annual meeting of National Association for Research in Science Teaching. Chicago, Illinois, April, 1972.

Roshal, S. M., Frieze, I., & Wood, J. T. A multitrait-multimethod validation of measures of student attitudes toward school learning and toward technology in sixth grade children. Educational and Psychological Measurement, 1971, 31, 999-106.

Scharf, E. S. The use of the semantic differential in measuring attitudes of elementary school children toward mathematics. School Science and Mathematics, 1971, 71, 641-649.

Silance, E. B., & Remmers, H. H. An experimental generalized master scale: a scale to measure attitude toward any school subject. Purdue University Studies of Higher Education, 1934, 35, 84-88.

Wrightsman, L. S., Nelson, R. H. & Taranto, M. The construction and validation of a scale to measure children's school morale. Paper presented at American Educational Research Association Convention, Chicago, February, 1968.

A PRIMER ON SAMPLING FOR STATEWIDE ASSESSMENT

Richard M. Jaeger

PREFACE

This paper is a brief introduction to finite population sampling methods, specially prepared for those concerned with statewide assessment programs. The sampling procedures described in the paper are those most likely to be useful in achieving the objectives of statewide assessment.

The paper is intentionally nonmathematical. While it presumes knowledge of the fundamental concepts of statistical inference, it does not require any prior exposure to the formalities of sampling. All sampling terms used in the paper are carefully defined. Descriptions of sampling procedures make use of these definitions and avoid unnecessary technicalities. The paper is intended to be a resource for those engaged in the practice of statewide assessment and makes no claim to comprehensiveness as a theoretical treatise.

Helpful suggestions and clarifications of some otherwise opaque issues were provided by Nancy Bruno, Paul Campbell, Henry Dyer and Robert Linn. I want to express my appreciation for their careful reviews of early drafts. I am solely responsible for any remaining inaccuracies.

Princeton, New Jersey

Richard M. Jaeger

Introduction

When a statewide assessment is planned, one of the first issues that arise is who should be tested? Even after a state has decided to test students in certain grades or at certain age levels, the question of who should be tested remains. Should all fourth graders be tested or should some be selected for testing?

In some states, the objectives and purposes that give rise to assessment include a desire to secure test results for each student in a grade; the assessment goals include individual assessment as well as institutional assessment. When individual assessment is desired, the who-to-test question is answered by the selection of a grade or age level for assessment. When individual measurement is not a goal of statewide assessment, it is usually economical and administratively desirable to select a sample of students for testing rather than testing all students.

This paper is intended to be a primer on sampling for statewide assessment. If its purpose is achieved, the careful reader will gain substantial knowledge about the promises and pitfalls of sampling for assessment. The reader will not become an instant sampling expert; no short paper can accomplish that goal. Instead, the dedicated reader will become a "sampling conversationalist" able to meet a sampling expert at least half way and able to knowledgeably discuss sampling issues important to his state's assessment. Further, he will be able to converse in the language of the expert.

The goal of creating "sampling conversationalists" will be pursued in three ways:

1) By defining terms and concepts basic to sampling theory and its applications
2) By illustrating some of the ways sampling procedures can be used to achieve realistic assessment objectives
3) By describing issues that arise when sampling procedures are used and the factors that contribute to their resolution

The balance of this paper is in two parts. The first part provides definitions of some of the most important terms and concepts fundamental to the language of sampling. In the second, consideration is given to two potential objectives of a statewide assessment and the ways various sampling procedures can contribute to their achievement. In part two, the reader is faced with alternatives and choices and then presented with facts to help him make decisions.

Some Terms and Concepts

Population: In any sampling study, there is a definable group or aggregation of elements from which samples are selected. This aggregation of elements is called the population of the study. Technically, any aggregation of elements that have at least one attribute in common can form a population. In a statewide assessment, some examples of populations that might be of interest are all public schools in the state that enroll sixth graders, all sixth graders enrolled in public schools in the state and all public-school sixth graders in the state who are children of migrant agricultural workers. From these examples, it is clear that populations can be composed of individuals or institutions. Similarly, populations can be composed of people or things. The first population--all public schools in the state that enroll sixth graders--is defined by two attributes: control of school (public) and grade-level offerings (sixth grade); the second population is also defined by two attributes: grade-level and public-school enrollment; the third population has three defining attributes: grade-level, public-school enrollment, and parental occupation.

These examples of populations have some important characteristics in common. Each is composed of a finite number of elements (sixth graders in the state, schools with sixth graders in the state, and so on), and each is defined by attributes that are easily recognized. That is, one can easily decide whether an element is or is not a member of the population.

Some populations that are infinite in size may be encountered in a statewide assessment. An example of an infinite population is "all multiple-choice test items that could ever be written, that purport to measure reading comprehension." In contrast to the first examples, this population is not defined by attributes that are easily recognized. If faced with a test item that contained a paragraph of prose followed by four questions on the main theme of the paragraph, most of us would say that the item was a "reading comprehension" item, and therefore a member of the population. But what about an arithmetic work problem--"If it took six men five days to dig a ditch..."? Clearly, reading comprehension is a skill required to answer the item correctly. Yet it requires more than reading comprehension to compute a correct solution. Is the item a member of the population? The answer is debatable.

All of the sampling procedures discussed in this paper assume that the populations to be sampled are finite. This is a realistic assumption whenever students, classes, schools or school districts are sampled. Unlike finite populations, infinite populations are somewhat intangible and exist only in the mind of the beholder. However, there is a well-developed theory of sampling from infinite populations, so they present no insurmountable statistical problems.

Another way of defining a population is "the aggregation of elements that is of central interest in a study." This is an admittedly loose definition that might upset some statistical purists, but it helps to point out the practical significance of populations. In a real-world study such as a statewide assessment, populations are not theoretically defined entities that exist for the fascination of statisticians; they are the central focus of the study. For example, in your statewide assessment, you may want to know the proportion of public-school fourth graders whose reading comprehension score is below the 25th percentile on a national norm distribution. Here, the population of interest is all fourth graders enrolled in the public schools of your state. The population is real and of practical interest. If you test every public school fourth grader in the state, you can determine the proportion exactly (provided there are no missing data, all absentees are tested at a later date, and so on).

Sampling unit: Populations are made up of elements termed sampling units. The sampling units into which the population is divided must be unique, in the sense that they do not overlap, and must, when aggregated, define the whole of the population of interest. Sampling units that might be used in statewide assessments include students, class sections, homerooms, teachers, schools, and school districts. These examples of sampling units clearly define unique elements (one student is different from another; schools that have the same grade levels are generally unique units) that can be readily counted and aggregated.

The definitions given for "population" and "sampling unit" may appear to be circular. But perhaps that's as it should be, since sampling units, when aggregated, make up a population, and a population is an aggregation of sampling units.

Sampling frame: When selecting a sample, one is, in fact, selecting sampling units from the aggregation that composes the population. For a unit to be selected, it must be identifiable. A list that uniquely identifies all of the units in a finite population is termed a sampling frame. A sampling frame for statewide assessment might consist of a list of all schools in the state that enroll pupils in grades one through six or a list of all secondary students enrolled full time in vocational education programs.

When assembling a sampling frame, care must be taken to ensure that it corresponds precisely to the population of interest. In the first example above, a sampling frame that consists of all schools in the state that enroll pupils in grades one through six would be composed of nonpublic schools as well as public schools. If the population of interest consisted only of public elementary schools, this sampling frame would be inappropriate. First, nonpublic schools would be listed in the frame although they are not elements of the population of interest. The erroneous listing of elements outside the population of interest is known as "overregistration." Second, the definition of an elementary school differs from state to state. In some states, a school is classified as an elementary school if it enrolls pupils in any grade between kindergarten and grade six. In other states, an elementary school is defined as a school that enrolls pupils in any grade between kindergarten and grade eight. In states with the latter definition, there may be schools that enroll only seventh and eighth graders that would be elements of a population of elementary schools. Yet these schools would be excluded from a sampling frame that listed schools with pupils in grades one through six. In this case, elements of the population of interest (all public elementary schools) would be excluded from the sampling frame (all schools that enroll pupils in grades one through six). This type of error in constructing a sampling frame is known as "underregistration."

The point to be made is that populations of interest in statewide assessment should be clearly and precisely defined. Then sampling frames that include only elements in the populations of interest and all elements in the populations of interest should be carefully constructed.

Probability sampling procedures: When sampling is used in statewide assessments, the financial objectives are clear. The desire is to save money and time by measuring or testing only a sample of students yet be able to make accurate statements about a population of students. Probability sampling procedures often allow these objectives to be achieved and, in addition, allow one to determine the likelihood of making inaccurate statements about a population.

Probability sampling procedures have three characteristics in common. First, the procedures are applied to populations where the units which compose the population and the units which are excluded from the population are explicitly defined. That is, given a potential sampling unit, one can say unequivocally whether it is in the population or not. Second, the chances (or probability) of selecting any potential sample can be specified. Third, every sampling unit in the population has a positive chance of being selected. It isn't necessary that every potential sample have an equal chance of being selected, just that the chance of selecting any potential sample can be specified.

The formal definition of a probability sampling procedure might appear somewhat formidable and perhaps unenlightening as well. Sometimes even simple things are obscured by formality (a square is a right parallelopiped composed of four pairwise orthogonal line segments). Instead of pursuing the definition further, consider some sampling methods that are not probability sampling procedures. Assume that an assessment objective is to determine the average social studies achievement of eighth graders in each school district in the state. Suppose that a particularly large school district decides to test eighth graders in half its schools and use their average achievement as an estimate of the average for all eighth graders. Suppose they decide to select for testing those schools that are closest to the district research office. With this plan, they'll select the school closest to the research office first, the second closest school second, and so on, until half the schools in the district have been "sampled." This isn't a probability sampling procedure, because it violates the third characteristic of such procedures. All the schools with eighth graders that are furthest from the district research office are contained in the sampling frame, but they don't have any chance (zero probability) of being selected. This same violation would occur with any sampling procedure that selects schools only from a prescribed section of the district.

These sampling procedures cause problems not because they violate an arbitrary rule, but because they are likely to produce samples that don't represent the population. The district research office is probably in the older or downtown area of the system. Schools near it are more likely to enroll students from lower socioeconomic status families than in the district as a whole, and the achievement of these students is therefore likely to be lower than in the district as a whole. So again, the rules are not just statistical artifacts. They help to prevent trouble in the practical world of assessment.

Estimate, population parameter and estimator: In addition to providing procedures for collecting data, sampling theory provides formulas for estimating characteristics of populations, such as averages, proportions, and totals. When a sample is drawn from a population, and a statistic (such as an average) is computed from data on the units sampled, the number that results is called an estimate. For example, if it is found that a sample of 10 students selected from a population of 200 has an average arithmetic score of 42, the number 42 is an estimate of the average for the entire population of 200. The average for the entire population would be an example of a population parameter. In general, population parameters are unknown characteristics of populations that survey researchers would like to know. If every element in a population is measured, the value of the population parameter can be determined. Instead of measuring every population element, a survey researcher will measure only elements in a sample and, from these data, compute an estimate of the population parameter. Formulas that are used to compute estimates from sample data are termed estimators.

In a statewide assessment, the average educational level of teachers in the state might be estimated by sending a questionnaire to a sample of teachers and computing an average for them. An average computed from the questionnaire responses of the sample is an estimate, and a formula used to compute the average for the sample of teachers in an estimator.

Estimator bias: When a population is finite, the number of different samples that can be drawn from it is also finite. A list can be made for any finite population containing all of the samples of a given size that could possibly be drawn from it. For example, suppose that a school district has four high schools and an assessment director wants to sample two of the four. If the schools are numbered from one to four, the six different samples of two schools that could be drawn are as follows:

Sample	Schools in Sample
A	1, 2
B	1, 3
C	1, 4
D	2, 3
E	2, 4
F	3, 4

Suppose the assessment director wants to know the average number of certified science teachers per high school in the district and decides to estimate the average by collecting data in two of the four schools. In this example, the population parameter is the actual average per school for the four schools in the district. Data from each sample would provide an estimate of this population parameter and, since six different samples could be selected, six different estimates are possible.

Continuing the example, suppose that an estimate of the population average per school was actually calculated using data from each sample, and the six estimates were then tabulated. It would then be possible to calculate the average of these six estimates. If the average value of the estimates were equal to the population average, the estimator (formula used to calculate each estimate) would be termed an unbiased estimator. If, on the other hand, the average of the sample estimates was either larger or smaller than the population average, the estimator would be biased.

In general, an estimator is said to be biased if the average of the estimates it would produce (if the average were to be taken over all possible samples of a given size) were either larger or smaller than the population parameter. If the average of all estimates were to equal the population parameter, the estimator would be termed unbiased.

It should be intuitively clear that unbiased estimators are desirable. An assessment director would be happiest if every estimate computed from a sample were equal to the population parameter of interest. Since this utopian condition will hardly ever be true, it is at least nice to have the average of the estimates equal the population parameter.

Although unbiased estimators are desirable, a biased estimator can sometimes be useful if the magnitude of the bias (the difference between the average estimate and the population parameter) is small. Under some conditions likely to be encountered in a statewide assessment, an unbiased estimator may actually be rejected in favor of a biased one.

At this point, the reader may wonder how estimator bias can be computed using data from a single sample. The answer is that it can't be computed from sample data. To compute bias, one would have to know the value of the population parameter. If the population parameter were known, there would be no reason to sample.

The bias (or lack of bias) of a sampling and estimation procedure is actually determined from the estimator used (a mathematical formula) and the mathematical assumptions that underlie the sampling procedure. Determination of bias is an algebraic procedure that doesn't depend upon data at all (Murthy, 1967; Cochran, 1963).

NUMERICAL EXAMPLE. Suppose that the average number of certified science teachers per school was known to be equal to 3.5 for the four schools in the district, and the estimates computed for the six possible samples were as follows:*

Sample	*Schools in Sample*		*Estimate*
A	1, *2*		*4.3*
B	1, *3*		*3.2*
C	1, *4*		*2.8*
D	*2, 3*		*3.7*
E	*2, 4*		*3.2*
F	*3, 4*		*3.9*
		Total	21.1

The average of the six estimates would equal

$$\frac{21.1}{6} = 3.52.$$

The estimator used would then be slightly biased since the true value of the population parameter is 3.50 and the average of the estimates produced by all possible samples of size two is 3.52. The magnitude of the bias is equal to the difference between the population parameter value, and the average of the six estimates: 3.50-3.52 = -0.02.

**In this numerical example and in those that follow, hypothetical data are used. It is critically important to recognize that these examples have been constructed solely to illustrate the definitions of sampling concepts presented in the main body of the paper. Each example assumes a situation that is totally fictitious and unlike the situations that will be encountered in practice. Namely, it is always assumed that the values of population parameters are known and that estimates are available for all of the samples that could possibly be selected.*

Variance, mean square error and efficiency: When an estimate of a population parameter is computed, it will rarely be equal to the population parameter. The difference between the estimate and the population parameter is known as an error of estimation. In the numerical example of the last section, the average number of certified science teachers per school was assumed to be equal to 3.5 for the four schools in the district, and the estimate computed from Sample F was assumed to be 3.9. With these assumptions, the error of estimate would be (3.5) - (3.9) or -0.4.

If an estimator is unbiased, its variance is equal to the average of the squared errors of estimate when the average is computed over all possible samples of a given size. Suppose that the estimator in the example of the last section had been unbiased; then, applying this formula for variance, the error of estimation would be computed for each of the six sample estimates, each of these would be squared, and the average of the six squared errors would equal the variance.

For a given sampling procedure and samples of a given size, the most desirable unbiased estimator is the one with the smallest variance. The smaller the variance of an unbiased estimator, the smaller the chance that a large estimation error can occur.

When an estimator is biased, its variance is also defined as the average of squares of differences. But instead of squaring the difference between each estimate and the population parameter, the variance of a biased estimator requires that the difference between each estimate and the average of all estimates be squared. The average of the squares of these differences is taken over all potential samples of a given size.

In a practical sampling situation, population parameters will not be known. (If they were known, sampling would be unnecessary). Additionally, only one sample will be selected, and only one estimate of the population parameter will be computed. The variance of the sample estimate (see the following section of the text) will not be directly computable from the data provided by a single sample. However, the variance of the sample estimate can almost always be estimated from the data provided by a single sample, and this estimate will almost always be computed in practice.

NUMERICAL EXAMPLE. Consider once again the hypothetical data presented in the last numerical example. In that example, the average number of certified science teachers per school was assumed to equal 3.5 in a school district with four schools. All possible samples of two schools were identified, and estimates of the average number of certified science teachers per school were assumed to be as follows:

Sample	*Estimate*
A	*4.3*
B	*3.2*
C	*2.8*
D	*3.7*
E	*3.2*
F	*3.9*

The average of these estimates was found to equal 3.52. These data may now be used to compute the variance of the estimator:

Sample	*Estimate*	*Difference Between Estimate and Average*	*Square of Difference*
A	*4.3*	*4.3-3.52 = 0.78*	*0.6084*
B	*3.2*	*3.2-3.52 = -0.32*	*0.1024*
C	*2.8*	*2.8-3.52 = -0.72*	*0.5184*
D	*3.7*	*3.7-3.52 = 0.18*	*0.0324*
E	*3.2*	*3.2-3.52 = 0.32*	*0.1024*
F	*3.9*	*3.9-3.52 = 0.38*	*0.1444*
		Sum of Squares:	*1.5084*

Variance of Estimator = (1.5084)/(6) = 0.2514

The definitions of variance for biased estimators and unbiased estimators are illustrated by Figures 1A and 1B on page 11. Each figure shows a distribution of estimates across all potential samples from a population. In Figure 1A, the average of all estimates and the population parameter have different values, and the difference between them is equal to the bias of the estimator. In Figure 1B, the average of all estimates and the population parameter have the same value, since the estimator is unbiased.

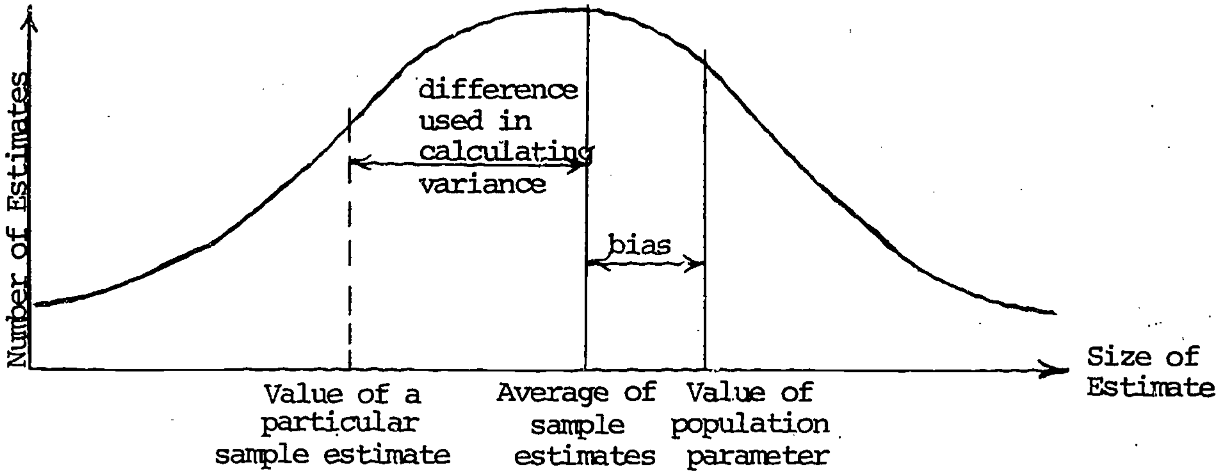

Number of Estimates
difference used in calculating variance
bias
Size of Estimate
Value of a particular sample estimate
Average of sample estimates
Value of population parameter

Figure 1A: Distribution of estimates for a biased estimator

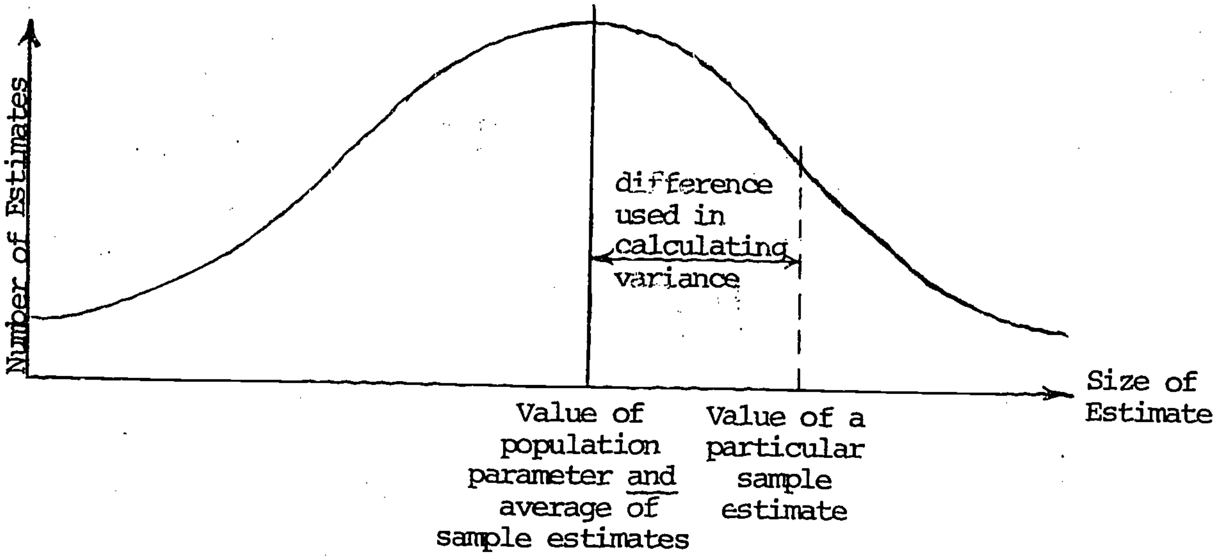

Number of Estimates
difference used in calculating variance
Size of Estimate
Value of population parameter and average of sample estimates
Value of a particular sample estimate

Figure 1B: Distribution of estimates for an unbiased estimator

If an assessment director has a choice of using two unbiased estimators, the one with the smallest variance should be selected. But what if the choice is between a biased estimator and an unbiased estimator? The biased estimator may have the smallest variance, but its bias may be large, and the proper choice is unclear. The assessment director needs some way of comparing the magnitude of estimation errors of biased and unbiased estimators. A useful measure for this purpose is called the mean square error. Mean square error equals the sum of the estimator variance and the square of the estimator bias.

$$\text{Mean square error} = \text{Variance} + (\text{Bias})^2.$$

NUMERICAL EXAMPLE: Using the data of the previous numerical examples in the formula for the mean square error,

$$\begin{aligned} \textit{Mean Square Error} &= 0.2514 + (-0.02)^2 \\ &= 0.2514 + 0.0004 \\ &= 0.2518 \end{aligned}$$

In this numerical example, the mean square error of the estimator is clearly dominated by the variance. Although the estimator is biased, the magnitude of the bias is very small, and bias contributes an insignificant amount to the mean square error.

For an unbiased estimator, the mean square error and the variance are equal, since the bias is zero.

For a given sample size, an estimator that has a smaller mean square error than another is said to be more efficient. For a given sampling procedure, the most efficient estimator should always be used, since it will provide the smallest estimation errors, on the average. When different sampling procedures are used, a less efficient estimator may be preferred if its sampling procedure is less costly or more convenient. In the practical world of statewide assessment, it may be worthwhile to take a larger sample if the sampling procedure that can be used is more administratively convenient or less expensive to complete.

Consistency: Some amount of error in the estimation of population parameters from sample data is almost inevitable. However, the magnitude of errors likely to occur can often be controlled. With some sampling and estimation procedures, the mean square error value can be reduced by drawing larger and larger samples, and estimation error is reduced to zero when the sample size equals the population size. Such procedures are said to provide consistent estimation. A sampling and estimation procedure is said to be inconsistent if sampling errors can occur even when the sample size equals the population size.

When lack of consistency is encountered in practice, the sampling is usually being done "with replacement." In a "with replacement" procedure, an element of a population can enter the same sample more than once. Although lack of consistency can occur when elements are sampled without replacement (once an element is sampled it is removed from the population), it is not encountered in practical problems.

As an example of a "with replacement" sampling procedure, consider the case discussed in conjunction with estimator bias above. In that example, two schools were sampled from a population of four schools. If sampling were to be done with replacement, 10 different samples of 2 schools could be drawn. In addition to the six samples listed in the previous example, the following are possibilities:

Sample	Schools in Sample
G	1, 1
H	2, 2
I	3, 3
J	4, 4

More to the point, one could select many different samples of four schools, such as:

Sample	Schools in Sample
A	1, 2, 3, 3
B	1, 1, 2, 3
C	1, 1, 3, 4
D	1, 1, 1, 1
E	3, 4, 4, 4

Unless the number of certified science teachers was the same in all schools, each of these samples would provide a different estimate of the average number of science teachers per school. As a result, sampling errors could occur even though the sample size and the population size were the same.

Lack of consistency becomes a problem of real concern in two situations. First, when the mean square error of an estimator is not reduced in size in some orderly way, as the sample size is made larger and larger. Second, when the size of the sample necessary to achieve an acceptable mean square error is close to the size of the population. Several sampling and estimation procedures that are otherwise attractive for statewide assessment may produce these problems in some situations. These procedures, and the potentially problematic conditions, are described in the next part of this paper.

NUMERICAL EXAMPLE. Consider once again the hypothetical situation described in previous numerical examples, but suppose that a "with replacement" sampling procedure is used. Assume that all samples of size one, two, three, and four schools are selected, and the mean square error of the estimator is computed for each sample size. Suppose that the results are as follows:

Sample Size	*Mean Square Error*
1	*1.25*
2	*0.64*
3	*0.88*
4	*0.22*

This example illustrates two kinds of inconsistency. First, the mean square error does not become progressively smaller as the sample size is increased; the mean square error for samples of three schools is larger than the mean square error for samples of two schools. Second, the mean square error is larger than zero for samples of four schools, even though there are only four schools in the population.

Clearly, the first kind of inconsistency is intolerable. A sampling researcher never knows how large the mean square error will be, although it can be estimated for many sampling procedures. Unless estimates are made for every possible sample size (which is sometimes impossible), the researcher can't determine an appropriate sample size with any degree of confidence; a large sample may be less efficient than a small sample.

Using Sampling in Statewide Assessment

Whether sampling is useful for statewide assessment depends primarily on the objectives of the assessment and secondarily on the capabilities of those conducting the assessment. For some assessment purposes (usually when assessment results are desired for individual students), sampling will not be useful at all. For other purposes, as when assessment results are desired for individual classrooms, sampling may be feasible but impractical. But much of the time, sampling will not only be feasible but a practical route to saving time, dollars and effort.

The capabilities of the agency conducting the assessment have been deemed secondary when considering the usefulness of sampling, since considerable help--through consultants or outside agencies--is likely to be readily available. Further, the costs of such assistance are likely to be more than repaid through the savings afforded by sampling.

Some sampling procedures are both feasible and practical for some assessment purposes, but infeasible or impractical for others. For example, simple random sampling (which is discussed below) may be impractical for determining the average achievement of pupils in a particular grade throughout a state (the impracticality stems from the need for a single list of all pupils enrolled throughout the state), but practical and feasible for determining the average achievement of pupils in a particular grade in each school in the state. In the latter case, separate simple random samples might be selected from each school using readily available lists in each school district.

To this point, this paper has been concerned with the language of sampling--basic terms and concepts necessary to an understanding of sampling and samplers. We shall now change course by considering two practical assessment objectives gleaned from actual state assessment reports, and describing how sampling procedures could be used in achieving these objectives.

Objective 1: Determining the Average Reading Achievement of All Fifth Graders in the State

An obvious way of determining the average reading achievement of all fifth grade pupils in a state is to test them all, record their scores, and compute the average. This procedure, known as taking a <u>census</u> of fifth graders, was actually followed in the state that reported this objective. For many objectives, and particularly when estimating statewide averages, taking a census is wasteful and unnecessary.

<u>Simple random sampling</u>: One procedure that could be used to achieve Objective 1 is called <u>simple random sampling</u>, a procedure in which every potential sample has an equal chance of being selected. Merely computing

the arithmetic average of data from a simple random sample will provide an estimate of the population average. This sampling and estimation procedure is unbiased and consistent, and there are well-known formulas for estimating the mean square error of the sample average (Hansen, Hurwitz and Madow, 1953).

To estimate the average reading achievement of fifth graders in a state through simple random sampling, the procedure would be as follows: First, a sampling frame would be constructed by listing each fifth grader enrolled in the state and assigning a unique number to each listed pupil. The sampling frame would include all enrolled fifth graders or only fifth graders enrolled in public schools, depending on the population of interest. Once the sampling frame was constructed, a table of random numbers would be used to select a sample of the desired size. A number would be drawn from the random number table and the pupil with the corresponding number would be added to the sample. If a number drawn from the table either exceeded the largest number on the list of pupils, or repeated a number already drawn, it would be discarded. Selection of random numbers from the table and corresponding pupils from the list would continue until the desired sample size was reached.

A practical problem that we have skirted so far will arise time and time again in sampling. Just what is the desired sample size and how can it be determined? With simple random sampling, the desired sample size can be computed through straightforward application of a formula given by Hansen, Hurwitz and Madow (1953), Cochran (1963) or in many other books on sampling. Rather than stating the formula here, we will consider some of the factors that enter into it. First of all, the size of a sample that's required to estimate a population parameter depends on the magnitude of the estimation errors that can be tolerated. The entire population must be sampled if the parameter must be known exactly. If a sample is taken, there will almost always be some estimation error, and for some samples the error may be very large. Since simple random sampling is consistent, the variance of estimation errors can be reduced by increasing the sample size.

Three factors enter the sample size formula for simple random sampling: the size of the population, the variance of the variable that is to be estimated, and the size of the estimation error that can be tolerated. Some rules of thumb for these factors are as follows: The larger the population size, the smaller the percentage that must be sampled in order to realize an estimator variance of a given size. For example, with a population of

100 pupils it might be necessary to sample 50 percent (or 50 out of 100); but with population of 10,000 pupils it might only be necessary to sample one percent (or 100 out of 10,000) to realize a given estimator variance. The larger the variance oi the variable for which a parameter is to be estimated, the larger the sample size required to achieve a given estimator variance. This is intuitively reasonable. If the variable (for Objective 1, reading achievement) has a large variance, estimates will fluctuate greatly from sample to sample; a larger sample size will be required to reduce its average fluctuations. Finally, the smaller the estimation error that can be tolerated, the larger will be the required sample size. Again, this rule is intuitively reasonable.

Should simple random sampling really be used to achieve Objective 1? Probably not, for the following reasons: First, there are other, more efficient sampling methods that can be used. Second, it would be administratively cumbersome to use simple random sampling. As previously mentioned, the assessment director would need a complete list of all fifth graders enrolled in the state. While such a list could probably be compiled in most states, its preexistence is doubtful, and its compilation would be expensive. When sampled fifth graders were actually tested, some classes of 25 would have 20 tested pupils, some would have only one or two tested pupils, and some would have none at all. Testing only some of the pupils in a classroom is administratively cumbersome and probably should be avoided unless the number of pupils drawn from each classroom is very small.

Simple random sampling is almost always discussed in sampling texts because it is a straightforward procedure and can be used to illustrate important sampling properties. It also provides a benchmark against which the efficiency of more sophisticated sampling procedures can be compared. For statewide assessment the practicality of simple random sampling is limited, although it may be useful when the objective is to estimate some property of schools or school districts.

Stratified random sampling: An alternative to simple random sampling that could be used to achieve Objective 1 is stratified random sampling, which is generally more efficient because it takes advantage of facts that are known about the elements of a population. Stratified random sampling can be contrasted with simple random sampling by considering a specific example. Suppose that the size of a simple random sample necessary to estimate the average reading achievement of a state's fifth graders was found to be 200. Following the procedure for selecting a simple random sample, it is possible that the 200 pupils selected might have an achievement average that was far higher than the average for all fifth graders in the state. This would almost surely be the case if most of the pupils in the sample had verbal IQ scores that were, say, above 130. Suppose it was possible to guard against samples that had almost all high-IQ pupils by ensuring that any sample selected would have some low-IQ pupils, some mid-IQ pupils, and some high-IQ pupils, with percentages of each similar to the percentages for the whole state. Samples of pupils that came close to representing the state's fifth graders on verbal IQ would probably do a good job of representing them on reading achievement. This is true because verbal IQ score and reading achievement are highly related; those with high verbal IQ scores are likely to have high reading achievement scores, and those with low verbal IQ scores are likely to have low reading achievement scores. Use of known relationships among variables and available data on sampling units is what makes stratified sampling efficient. Stratified sampling prevents the selection of extremely unrepresentative samples (such as all high-IQ pupils) and thereby prevents large estimation errors. To achieve an estimator variance of a given size, stratified sampling will therefore require a smaller sample size than will simple random sampling.

In stratified random sampling, elements of the population are first classified into categories called *strata* according to their values on one or more *stratification variables*. In the previous example, verbal IQ played the role of a stratification variable. Any variable for which a value is known for every element of the population can be used as a stratification variable. However, stratified sampling won't be efficient unless the stratification variable and the variable for which estimates are desired (reading achievement in the previous example) are highly related.

Considering the previous example more explicitly, suppose that verbal IQ was to be used as a stratification variable and the parameter to be estimated was the average reading achievement of all fifth graders in a state. The first step in using stratified random sampling would be to define appropriate strata. For example, low-IQ pupils might be defined as those with verbal IQ scores below 85, mid-IQ pupils might be defined as those with verbal IQ scores between 86 and 115, and high-IQ pupils as those with verbal IQ scores of 116 or more. These IQ intervals would define three strata and might be labeled stratum 1, stratum 2 and stratum 3. Once the strata were defined, each fifth grader in the state would be classified as a member of stratum 1, 2 or 3 depending on his (her) verbal IQ score. When all fifth graders in the state had been assigned to strata, a simple random sample of pupils would be drawn from each stratum. The average reading achievement of pupils sampled from each stratum would then be calculated, and these averages would be weighted appropriately to form an estimate of the average achievement of fifth graders throughout the state. The estimator would be both unbiased and consistent.

For estimating a statewide average, stratified random sampling has the same disadvantages as simple random sampling. It requires a sampling frame that lists all fifth graders in the state. In addition, it might result in selection of a few pupils from some classes and many pupils from others. It thus has the potential of being administratively disruptive in some schools and districts.

The main advantage of stratified random sampling is its efficiency (when the right stratification variables are used). In addition, when stratified sampling is used in statewide assessment or in other educational data-collection programs, the information needed for stratification is generally available. During the last decade at least, group IQ testing has been almost universal, and nearly all school districts administer standardized achievement tests (Goslin, 1967). In addition, school systems record all manner of information on their pupils such as parental occupations, educational levels of parents, and sizes of pupils' families. All of these variables tend to be highly related to current educational achievement (Mollenkopf and Melville, 1956; Burkhead, 1967) and if available would be quite useful as stratification variables in statewide assessment.

In theory, strata can be defined by any number of variables. One could, for example, stratify pupils by IQ scores and status level of father's occupation. The strata thus formed might be labeled low-IQ and low-status occupation, low-IQ and mid-status occupation, low-IQ and high-status occupation, mid-IQ and low-status occupation, etc. Stratification by two or more variables is only efficient when each stratification variable is highly related to the variable for which estimates are sought and when the stratification variables are not highly related among themselves. The previous example, stratification of pupils by IQ level and by status level of father's occupation, would probably be an unnecessarily cumbersome procedure. Although reading achievement is highly related to both IQ level and status-level of father's occupation, the two stratification variables are themselves highly related. Pupils from high-status homes tend to have higher IQ levels, and vice versa. Stratifying pupils by these two variables is therefore redundant; stratification by either variable would be almost as efficient as stratification by both, although IQ level would probably be a better stratification variable than would father's occupation.

Practical use of stratified sampling requires several design decisions in addition to those already discussed. Once stratification variables have been chosen, the sample designer must decide how many strata to use, the limits or boundaries for each stratum (e.g., IQ below 90, IQ between 91-110, and so on), the size of the sample to select, and the number of units to sample from each stratum. Each of these topics has been the subject of theoretical and empirical study in the theory of sampling. Again, some practical factors that influence the decisions will be described. The choice of number of strata depends on the magnitude of the relationship between the stratification variable and the variable for which estimates are sought. The stronger the relationship, the larger the number of strata that will prove useful, although practical limits are reached very quickly. Even when the stratification variable and the variable of interest have a correlation coefficient of 0.90, there is not much advantage to using more than four strata (Cochran, 1963). The problem of determining boundaries for strata so as to make stratified sampling as efficient as possible has been given considerable attention by Dalenius and Hodges (1959). They provide formulas that can be used in practice but defy simple, intuitive explanation. Explicit formulas also exist for determining the sample size to use in stratified sampling. As in simple random sampling, required sample size depends on the population size and the size of the estimation errors one can tolerate. Unlike simple random sampling, the sample size for stratified sampling also depends on how well the population has been stratified. The object of stratification is to form categories within which sampling units are as nearly alike as possible on the variable of interest. The more nearly this has been accomplished, the smaller will be the sample size required to achieve a given estimator variance. Determination of the number of units to be sampled from each stratum is generally handled in one of two ways. Using a procedure termed optimal allocation, a specific formula indicates the sample size for each stratum. The advantage as efficient as possible (hence the term optimal). An alternative procedure is termed proportional allocation. With proportional allocation, the size of the sample selected from each stratum is proportional to the number of population elements in the stratum. The advantages of proportional allocation include simplified estimation formulas and assurance that the stratified sampling procedure will be at least as efficient as simple random sampling.

Systematic Sampling: The average reading achievement of fifth graders in a state could also be estimated by using a systematic sampling procedure. Several systematic sampling procedures have been developed in the last two decades, but only the one used most widely--linear systematic sampling--will be considered.

Like simple random sampling, linear systematic sampling would require a sampling frame of fifth-grade pupils. Instead of consulting a table of random numbers to determine each sampled pupil, a random number table is consulted only once with linear systematic sampling. The sampling frame of pupils is considered to be an ordered list. The first sampled pupil is selected randomly, and successive pupils are selected at multiples of a constant interval beyond the first. A specific example may help to clarify the procedure.

Suppose it was desired to select a linear systematic sample consisting of ten percent of the fifth graders in the population. To determine the first sampled pupil, a number between one and ten would be drawn from a random-number table. The pupil with the corresponding number on the sampling frame would become the first sampled pupil. Thereafter, every tenth pupil would be sampled. Thus, if the random number six were drawn from the table, the first sampled pupil would be the one listed sixth in the frame, the next sampled pupil would be listed 16th in the frame, the next 26th, and so on, until the sampling frame had been exhausted.

> *NUMERICAL EXAMPLE. Consider the selection of ten percent systematic sample from a population of fifth-grade pupils. Suppose that a table of random numbers had been consulted to select a number between one and ten, and that the number drawn was six. If the sampling frame were as follows, the sampled pupils would be those marked with an asterisk:*

Pupil Number	Pupil Name
1	Murphy, John
2	Centra, Paul
3	Bruno, Barbara
4	Aron, Carol
5	Parker, Mary
*6	Nesbitt, William
7	Lee, Marjorie
8	Sinclair, Susan
9	Thomas, George
10	Wichert, Jane
11	Urban, Paula
12	Mann, Marcia
13	Tocco, Brenda
14	Malcolm, Thomas
15	Angoff, Douglas
*16	Fouratt, Sharron
17	Brambley, Joan
18	Willis, Kevin
19	Picard, Ronald
20	Libby, Linda
21	Arcieri, Sheryl
22	Kristof, Charles
23	Patterson, Virginia
24	Johnson, Elmer
25	Saxe, Anne
*26	Stahl, Mildred
27	Walsh, Helen
28	Adams, Patricia

.

.

.

The three dots signify the continuation of the list and the selection of every tenth pupil beyond the 26th until the entire sampling frame had been exhausted. Thus if the list contained 240 pupils, the last one selected for the sample would be pupil number 236.

Systematic sampling has the advantage that it is easy to apply by hand, whereas simple random sampling or stratified random sampling are quite tedious without a computer when a sample of appreciable size must be drawn. When used in an assessment program, systematic sampling would also ensure that the numbers of pupils sampled from each classroom were approximately equal provided the sampling frame listed pupils sequentially by classroom. Like simple random sampling, though, systematic sampling would require a list of all fifth graders in the state.

Unlike simple random sampling and stratified sampling, linear systematic sampling is sometimes undependable. It is not always consistent, and there are no really good ways to estimate mean square error. Conversely, linear systematic sampling can be very efficient if the list used for sampling is carefully constructed. If pupils were listed alphabetically in the sampling frame, one would suppose that their average achievement might be estimated about as efficiently as with simple random sampling. In fact, alphabetic listing of pupils sometimes results in more efficient estimation (Jaeger, 1970), although this won't always be the case. Real gains in the efficiency of systematic sampling can be realized by listing pupils in increasing order on some variable that is highly related to the variable of interest. For example, if a linear systematic sample of fifth graders was selected from a sampling frame in which pupils were listed in increasing order of their verbal IQ scores, average reading achievement could be estimated very efficiently. The effect of such ordered listings is much the same as the effect of stratification since sampling from an ordered list ensures that some pupils are sampled at all levels of the variable used for ordering.

Linear systematic sampling is one of those procedures mentioned earlier that isn't always consistent and, depending on the relationship between the sample size and the population size, may lead to biased estimation. Usually the magnitude of the estimation bias is inconsequential, but the lack of consistency may prove to be a serious problem. If sampling must be done without a computer and if the required sample size is large, linear systematic sampling should be considered for statewide assessment. Otherwise, alternative sampling procedures (such as stratified sampling) will provide more dependable results.

Cluster Sampling: In the sampling procedures discussed to this point, the sampling units used were basic elements of a population, e.g., individual pupils. In cluster sampling, the sampling units are not basic population elements but are groups or aggregations of such elements. These groups of elements are termed clusters.

In most applications of cluster sampling, the clusters used are naturally occurring groups. In surveys of consumer behavior, for example, homes are frequently used as sampling units. When estimating the average achievement of fifth graders throughout a state, several naturally occurring clusters of pupils might be used--school districts, schools, or homerooms. Of course, these aren't the only possibilities for clusters. One might consider groups of students living in particular areas of the state or groups of pupils with last names beginning with the same letter. However, naturally occurring clusters afford far greater administrative convenience than would these contrived clusters. Pupils can readily be identified by classroom, school, or school district and could easily be assembled for testing and measurement on a homeroom-by-homeroom or school-by-school basis.

If a cluster sampling procedure is identified by the units used as clusters--school districts, schools, homerooms, or combinations of these--many different cluster sampling procedures could be used to gather data for Objective 1. Before enumerating some of the possibilities, let's consider one in detail and thereby introduce some of the language of cluster sampling.

Suppose it was decided to use schools as clusters and to test the reading achievement of all fifth graders enrolled in sampled schools. This procedure is an example of single-stage cluster sampling. The sampling plan would be carried out by first constructing a sampling frame of all schools in the state that enrolled fifth-grade pupils. A simple random sample of schools could then be selected, using a table of random numbers, just as in simple random sampling of pupils described above. All of the fifth-grade pupils in sampled schools would then be given a reading achievement test, and appropriate formulas would be applied to the test results in order to estimate average achievement for the state. The formulas to be used (estimators) are well known in the sampling theory literature and can be found in any standard text such as Cochran (1963).

This cluster-sampling procedure has some obvious administrative advantages. First, the state department of education is likely to have a complete list of schools that enroll fifth graders, although it probably doesn't have a list of fifth graders enrolled in the state. Thus, a ready-made sampling frame is likely to exist for this sampling procedure. Second, only a sample of schools will be involved in testing. Disruption of normal academic procedures will be confined to the sample of schools, the costs of distributing testing materials will be reduced, and administrative procedures will be simplified.

The administrative convenience of this sampling procedure is likely to be offset by a substantial reduction in efficiency. In almost all cases, cluster sampling of schools will be far less efficient than simple random sampling of pupils. The "almost" is inserted in the previous sentence because there are notable exceptions to them. The efficiency of single-stage cluster sampling depends on many factors, some of which can be controlled by the sample designer. The composition of the clusters used influences efficiency to a large degree. Two extreme cases will illustrate this point. To take one extreme, suppose that all of the fifth graders in any given school had the same reading achievement score. In this case, testing all the fifth graders in a school would be a waste of time and money; the average achievement in a school could be determined by testing just one fifth grader. More to the point, the effective sample size is equal to the number of schools in which testing takes place rather than the number of pupils tested (since testing more than one pupil in a school would provide only redundant information). In technical terms, this extreme case represents a situation in which all of the elements within a cluster are completely homogeneous on the variable to be estimated. The other extreme would occur in a situation where the average reading achievement of fifth graders in each school was identical and equalled the average for the whole state. In this case, the average for the state could be estimated perfectly by collecting data in only one school since testing pupils in more than one school would provide only redundant information. In technical terms, this extreme represents a situation in which elements within a cluster are as heterogeneous as elements within the entire population and in which clusters are completely homogeneous. In real life, the composition of the population will fall somewhere between these extremes. For cluster sampling to be efficient, we would like the composition of the population to be similar to the second extreme: not much difference among clusters on the variable to be estimated and a lot of heterogeneity among elements in the same cluster. With this composition, only a few clusters need be sampled in order to get a good representation of the entire population.

Unfortunately, the naturally occurring clusters available for statewide assessments tend to provide homogeneity within clusters and heterogeneity between clusters for many variables likely to be of interest. Consider sampling of schools to estimate pupil achievement. At least before bussing for purposes of desegregation, the attendance areas of schools tended to be defined by neighborhoods that were relatively homogeneous in their socioeconomic and racial compositions. In a society where neighborhoods tend to be defined by people of the same social and economic level, it is natural that schools tend to be homogeneous in these variables. Since pupils' scores on achievement tests are highly related to the socioeconomic status of their families, schools also tend to be homogeneous in measured academic achievement.

The composition of the population of interest (e.g., all fifth graders in a state) is a factor beyond the control of the sample designer; whatever is found must be tolerated. However, there are factors that the user of cluster sampling can control so as to greatly increase sampling efficiency. One such factor is the estimation procedure employed. When the clusters to be sampled are not only heterogeneous but also tend to vary greatly in size (both are tendencies of schools and school districts), simple random sampling of clusters with unbiased estimation of averages is very inefficient. A more efficient alternative involves simple random sampling of clusters and use of an estimation procedure known as ratio estimation. To use ratio estimation, the number of elements in each cluster must be known--a requirement that is easily met in most assessment applications. The ratio estimator is biased but consistent. The amount of bias is likely to be small for populations used in statewide assessments, and the mean square error will usually be much smaller than that of the unbiased estimator. Formulas for ratio estimation can be found in Murthy (1967), Cochran (1963), and Hansen, Hurwitz and Madow (1953).

Additional alternatives modify both the sampling procedure and the estimation procedure used with single-stage cluster sampling. By definition, each cluster has an equal chance of being selected when clusters are sampled randomly. One alternative procedure, known as PPS sampling, selects clusters with probabilities proportional to their sizes. If schools were being used as clusters in order to estimate average fifth-grade reading achievement, the probability of selecting a given school would depend on its fifth-grade enrollment. A school with 200 fifth-graders would be twice as likely to enter the sample as would a school with 100 fifth graders. The PPS procedure provides not only a sampling method but associated estimators of averages, proportions and variances as well. It is simplest to do PPS sampling "with replacement" since selection probabilities vary as the sample is drawn when sampling is done without replacement. PPS sampling with replacement provides unbiased estimation but is an inconsistent procedure. The mean square error of the estimator gets consistently smaller as sample size is increased but does not go to zero when the sample size equals the population size. In practical situations, this lack of consistency will be a problem only when the required sample size is very close to the population size.

PPS sampling is efficient only when cluster size is highly related to the variable for which estimates are desired. Since school size and school-district size are not highly related to basic-skills achievement (Burkhead, 1967), PPS sampling will not be efficient for estimation of average achievement in a state. Some school and district "input" variables (such as the average value of the taxable property in an attendance area or district) are highly related to school or district size, and PPS sampling would probably be very efficient for estimation of these variables.

A final alternative, PPES sampling, is likely to be a very efficient way of estimating average achievement in a state. PPES stands for "probability proportional to expected size" (Cochran, 1963), a term that is appropriate in some sampling contexts but not in the context of statewide assessment. PPES sampling was first introduced to handle situations in which cluster sizes were not known exactly. In these cases, "expected sizes" rather than actual sizes were used.

In assessment applications, cluster sizes are usually known but are often nearly unrelated to the variables for which estimates are desired. The greater the relationship between the variable for which estimates are sought and the "expected size" variable, the higher the efficiency of PPES sampling. For this reason, clusters can be sampled with probabilities proportional to any variable that has a known value for every cluster in the population; the variable used can be totally unrelated to cluster size. Consider the case of Objective 1. Suppose that a group IQ test had been administered to every fourth grader in the state in the year preceding the current assessment. If the state had records containing the average IQ of fourth graders for each school and the fourth-grade enrollment of each school, the product of these two could be used very effectively as an "expected size" measure when estimating average fifth-grade reading achievement. This procedure would be highly efficient because the average of fourth-grade IQ scores and the average of fifth-grade reading achievement scores would be highly related across schools.

Like PPS sampling, PPES sampling results in unbiased but inconsistent estimation. Again, inconsistency will be a practical problem only when the required sample size is very close to the population size. Additional information on PPS sampling and PPES sampling can be found in Murthy (1967) and in Cochran (1963).

Instead of using schools as clusters, the average reading achievement of fifth graders in the state could be estimated by using either homerooms or school districts as clusters. Either of these single-stage cluster sampling procedures would be feasible provided appropriate sampling frames could be constructed. Undoubtedly, every state department of education has a complete listing of school districts that enroll fifth graders. A sampling frame of homerooms probably wouldn't exist in most states though, and sampling by homerooms would require a specially constructed frame. The cost of constructing a sampling frame of homerooms would probably be more than offset by the increased efficiency of a single-stage cluster sampling plan with homerooms as clusters. In most states, cluster sampling of homerooms would be far more efficient than cluster sampling of schools, and cluster sampling of schools would be more efficient than cluster sampling of districts. The increased efficiency is due in part to substantially greater size variability among districts than among schools, and among schools than among homerooms.

Thus far, we have considered only single-stage cluster sampling procedures. Many <u>multistage cluster sampling</u> procedures could be used to estimate the average reading achievement of a state's fifth graders. Possibilities include the following: 1) A random sample of schools could be drawn, and within sampled schools random samples of homerooms could be selected. All fifth graders in sampled homerooms would be tested. 2) A random sample of districts could be drawn, and within sampled districts random samples of schools could be selected. All fifth graders in sampled schools would be tested. 3) A random sample of districts could be drawn, and within sampled districts a random sample of homerooms could be selected. All fifth graders within sampled homerooms would be tested. 4) A random sample of districts could be selected, and within sampled districts random samples of fifth graders could be selected and tested. 5) A random sample of schools could be drawn, and within sampled schools random samples of fifth graders could be selected and tested. 6) A random sample of fifth-grade homerooms could be selected, and within sampled homerooms random samples of pupils could be drawn and tested. 7) A random sample of districts could be selected, random samples of schools could be drawn within sampled districts, and random samples of homerooms could be selected within each sampled school. All fifth-grade pupils within sampled homerooms would be tested. 8) A random sample of districts could be selected, random samples of schools could be drawn within sampled districts, random samples of homerooms could be drawn within sampled schools, and random samples of pupils would be selected and tested within sampled homerooms. Although these eight procedures do not exhaust the possibilities, they provide sufficient illustrations of the flexibility of cluster sampling.

Procedures 1 through 6 are examples of two-stage cluster sampling. In procedure 2, for example, sampling of districts constitutes the first stage (districts are termed primary sampling units or PSU's), and sampling of schools is the second stage. Schools would be called secondary sampling units. Procedure 7 is an example of a three-stage cluster sampling procedure with districts as PSU's, schools as secondary sampling units, and homerooms as tertiary sampling units. Procedure 8 is a four-stage cluster sampling procedure.

Multistage cluster sampling will often be more statistically efficient than single-stage cluster sampling. That is, the mean square error of the estimator will be smaller for a given number of elementary units in the sample. There are also some administrative advantages to multistage sampling. If sampling frames don't exist, they need only be constructed for a sample of PSU's. For example, if a state wanted to use homerooms as clusters but didn't have the required sampling frame, it could use two-stage sampling with districts as PSU's and homerooms as secondary sampling units. The district sample would be chosen first, and sampling frames of homerooms would be needed only for sampled districts.

Cluster sampling can also be used in combination with other procedures such as stratified sampling or systematic sampling. One could, for example, select samples of schools stratified by the average IQ level of enrolled fifth graders or by a measure of the average socioeconomic status of pupils' families. As another alternative, one could select a simple random sample of school districts and select systematic samples of fifth graders from lists arranged in order of increasing IQ scores within each sampled district. Each of these alternatives would be more efficient than multistage random sampling.

The final choice among cluster sampling procedures depends on many factors, not the least of which is previous knowledge of the population of interest. To choose among sampling procedures intelligently, one should have some idea of the degree of homogeneity within and among potential clusters and the relationships among variables for which estimates are sought and those that might be used for stratification or as measures of size. Even with these kinds of data, assurance that one has chosen the best of the available alternatives can only come through careful analysis and often lengthy computation. (See Appendix A).

It cannot be overemphasized that data typically available in schools and school districts can be used very effectively to design efficient sampling procedures. A wealth of information on students, teachers, classes, schools and school districts is routinely recorded and filed in school district offices and in offices of state departments of education. Data from previous testing programs are abundantly available in almost all school districts and states. Background information on pupils and teachers is also on file in most school districts. If judiciously selected and evaluated, these data can be used for stratification, for arrangement of populations in ordered lists, and for pre-testing of potentially efficient sampling procedures. This mechanical use of information to arrange and sort populations should not provoke charges of invasion of privacy since individuals' names need be associated with individual data elements only for purposes of sampling.

Matrix Sampling: Each of the sampling procedures considered to this point has assumed that all sampled pupils respond to the same set of measures--e.g., the same reading comprehension test. In the past ten years, researchers have paid increasing attention to procedures that sample test items as well as students. These procedures are termed multiple matrix sampling, and have been used successfully in National Assessment as well as in several statewide assessments.

Multiple matrix sampling could be used to estimate the average reading achievement of all fifth graders in a state. The procedure might be as follows. Suppose that a 50-item reading achievement test was to be used. Instead of administering the entire test to all sampled pupils, the test could be divided into five forms with ten items each. Each sampled pupil would then take a 10-item form instead of the entire 50-item test. Each of the 50 items would be used in a 10-item form, and approximately equal numbers of pupils would complete each 10-item form. Lord (1955; 1962) has developed formulas for estimating the average score pupils would have earned if each had completed the entire 50-item test. Empirical studies of the best way to divide tests into forms and the sizes of pupil samples to use with each form have been completed by Shoemaker (1970; 1971), Knapp (1968) and others.

To date, statistical procedures for analysis of multiple matrix sampling have been developed only for simple random sampling of items and pupils. Although more complex designs can be used, needed analytic procedures are not yet available.

Objective 2: Estimating the Proportion of Third Graders in Each School District who can Successfully Achieve an Arithmetic Objective

Some statewide assessments use test items that are specifically designed to measure the achievement of particular objectives. For example, an assessment might include items designed to measure achievement of the arithmetic objective, "addition of pairs of single-digit integers." Five such items might be administered to a pupil, and the pupil might be said to have achieved the objective provided he can successfully complete three of the five items.

Suppose that a statewide assessment contained such objectives-related items and that the principal purpose of the assessment was to determine the proportion of pupils in each of the state's school districts that had achieved each designated objective.

Many of the sampling procedures described above could be used to achieve Objective 2. Only in very small school districts (e.g., those with grade three enrollments under 200) would sampling be uneconomical. Among the procedures that might be used to achieve Objective 2 are simple random sampling of pupils, stratified random sampling, linear systematic sampling, and some forms of cluster sampling.

With Objective 2, each school district's third graders would constitute a separate population, and sampling in each school district could be handled differently; that is, one district might use simple random sampling, while another might use two-stage cluster sampling of schools and homerooms, with homerooms stratified by average ability level of pupils. In practice, use of several different sampling procedures would make good sense if the districts varied greatly in size. While cluster sampling would be infeasible in a small school district (one with only three elementary schools, for example), it might prove to be highly efficient in a state's largest school districts.

To accomplish Objective 2, simple random sampling would be handled just as it is described for Objective 1. Standard formulas exist for the estimation of proportions through simple random sampling just as they do for the estimation of mean square errors (Murthy, 1967; Hansen, Hurwitz and Madow, 1953).

When the objective is estimation of a proportion, stratified sampling is unlikely to afford appreciable increases in efficiency over simple random sampling. To be efficient, stratified sampling requires that variances within strata be much smaller than the variance within the whole population. The variances of proportions are very similar unless the proportions are extremely large or extremely small. (The variances of proportions in the range 0.2 to 0.8 are very similar.) Thus, little reduction in the variance of proportions can be gained from stratification.

Use of linear systematic sampling is just as reasonable for the achievement of Objective 2 as it was for the achievement of Objective 1. The same potential advantages and the same cautions apply. A school district is more likely than a state department of education to have past test data and other information on individual students. This information can be used to create ordered sampling frames, permitting systematic sampling from an ordered list.

Unless a school district is very large, multistage cluster sampling will not be practical. For moderately large school systems (enrollments of ten-thousand to thirty-thousand), single-stage cluster sampling of homerooms is likely to be administratively practical and statistically efficient for estimation of averages or proportions. Compiling a list of third-grade homerooms should not be difficult in a district of moderate size. Sampling by homeroom would permit testing of intact groups of pupils and would provide a convenient route for distribution of materials and handling of assessment materials in the field.

Multiple matrix sampling could also be economical and convenient in all but the smallest school systems. Shoemaker (1970) has shown that multiple matrix sampling is useful for estimation of average provided the population is no smaller than 300.

Summary

This paper was intended to help the reader become conversant with important sampling terms and concepts and to become aware of sampling procedures that might be used in a statewide assessment. It was not intended to create instant sample-design experts or sampling theorists.

If the reader has gained a basic understanding of such terms and concepts as estimate, estimator, population parameter, estimator bias, and so forth, and if some of the sampling options available for statewide assessments are now intelligible, the paper has accomplished its purpose.

Designing an efficient sample requires knowledge of the science of sampling. But perhaps more than in other statistically-oriented disciplines, good sample design is an art. It requires a sensitivity to the nature of the populations of interest and attention to information and data that the novice might feel is unrelated to the sampling task at hand. For these reasons, there is no substitute for experience when a truly efficient sample design is desired. Investment in expert sampling consultation will usually be repaid many times over by the economies an efficient design provides. But it behooves the assessment directors to be conversant, if not expert, on sampling and its potentials. If they know a little about the subject, the right questions can be asked and the right data can be provided The task of the sample designer will be made easier and the resulting product all the better.

REFERENCES

Burkhead, J. Input and output in large city high schools. New York: Syracuse Univer. Press, 1967.

Cochran, W. G. Sampling techniques. New York: John Wiley and Sons, 1963.

Dalenius, T. and Hodges, J. L. Jr. Minimum variance stratification. Journal of the American Statistical Association, 1959, 54, 88-101.

Goslin, D. Teachers and testing. New York: Russell Sage Foundation, 1967.

Hansen, M., Hurwitz, W., and Madow. W. G. Sample Survey Methods and Theory. 2 vols. New York: John Wiley and Sons, 1953.

Jaeger, R. M. Designing school testing programs for institutional appraisal: an application of sampling theory. Unpublished doctoral dissertation, Stanford Univer., 1970.

Knapp, Thomas R. An application of balanced incomplete block designs to the estimation of test norms. Educational and Psychological Measurement, 1968, 28, 265-272.

Lord, F. M. Sampling fluctuations resulting from the sampling of test items. Psychometrika, 1955, 20, 1-23.

Lord, F. M. Estimating norms by item sampling. Educational and Psychological Measurement, 1962, 22, 259-267.

Mollenkopf, W. G. and Melville, S. D. A study of secondary school characteristics as related to test scores. Research Bulletin 56-6. Princeton, N.J.: Educational Testing Service, 1956.

Murthy, M. N. Sampling theory and methods. Calcutta: Statistical Publishing Society, 1967.

Shoemaker, D. M. Allocation of items and examinees in estimating a norm distribution by item-sampling. Journal of Educational Measurement, 1970, 7, 123-128.

Shoemaker, D. M. Further results on the standard errors of estimate associated with item-examinee sampling procedures. Journal of Educational Measurement, 1971, 8, 215-220.

APPENDIX A

Evaluation of Alternative Cluster Sampling Procedures--An Example

The kinds of theoretical notions discussed in this paper (A procedure will be more efficient when cluster sizes don't vary much; heterogeneity within clusters and homogeneity between clusters will provide increased efficiency, and so forth.) provide some guidance for choosing among alternative cluster sampling procedures. In a specific application, assurance that one is using the best procedure can also be gained through analysis of data from the school district or state in which sampling is to be used.

Many characteristics of schools, school districts, and groups of students show remarkable stability from year to year. For example, the average basic skills achievement of a school's fourth-grade class is likely to be very similar in two successive years, as is the socioeconomic composition of the school's student body. When searching for a sampling procedure that provides maximum efficiency, one can take advantage of this kind of stability. The method is as follows: Use data from the previous school year to evaluate the efficiency of the sampling procedures being considered for the current year. Since it is unlikely that sampling has been used in the past, data will be available for all students, classes, and schools in the district or state. With data available for the entire population (a situation that will not hold for the current school year if sampling is used), results of sampling the previous year's population using a variety of procedures can be readily compared.

An example of this kind of evaluation uses data from a single school district called Anydistrict (Jaeger, 1970). For simplicity, computation of estimates and estimator variances will not be shown; only initial data and final results will be presented.

The population parameter to be estimated in this example is the average reading achievement of the district's sixth graders. The sixth-grade enrollment of the district is 1180, with 45 sixth-grade classes in 21 schools. Data available from the previous school year include the average sixth-grade reading achievement in each school, the sixth-grade enrollment in each school, and the average verbal ability score of fifth graders in each school. These data will be used to evaluate four alternative cluster sampling and estimation procedures: simple random sampling of schools with unbiased estimation; simple random sampling of schools with ratio estimation; sampling of schools with probabilities proportional to their sixth-grade enrollments (PPS sampling and estimation); and sampling of schools with probabilities proportional to totals of fifth-grade ability test scores (PPES sampling and estimation).

The evaluation of each cluster sampling procedure will use data from the entire population of 21 schools. With these data, estimator variances can be calculated exactly. It must be emphasized that data for the entire population will be available only when all sixth graders in the district are tested--a situation that will not obtain in the current school year, when sampling is used. The method, then, is to use population data from a previous school year to evaluate alternative sampling procedures and to assume that the most efficient procedure for one school year will also be most efficient for the next year. The assumption is generally sound.

The following table shows sixth-grade average reading achievement scores, sixth-grade enrollments, and average fifth-grade ability test scores for the 21 schools in the district under study. The data are real. They were provided by the research office of a medium-sized school district.

Table A: Sixth-Grade Average Reading Achievements, Sixth-Grade Enrollments, and Average Fifth-Grade Ability Test Scores for Elementary Schools in Anydistrict.

School Number	Average Grade 6 Reading Achievement*	Grade 6 Enrollment	Average Grade 5 Ability Score
1	66.11	56	33.54
2	66.83	65	32.96
3	71.27	71	38.06
4	56.09	58	33.81
5	64.57	47	34.29
6	71.09	66	37.84
7	74.89	55	36.70
8	70.67	99	37.69
9	74.51	57	39.06
10	68.13	40	37.19
11	70.02	59	36.10
12	72.57	72	39.90
13	58.86	43	35.36

*average number of test items correct

Table A: (continued)

School Number	Average Grade 6 Reading Achievement	Grade 6 Enrollment	Average Grade 5 Ability Score
14	66.35	63	36.20
15	70.71	38	36.92
16	65.82	51	34.42
17	70.98	51	35.15
18	67.56	41	33.51
19	82.21	29	40.76
20	65.61	74	35.02
21	51.14	49	30.18

The data in Table A were used in formulas for the variance of the estimated mean appropriate to each of the four cluster sampling and estimation procedures. In all cases, it was assumed that 10 of the 21 schools in Anydistrict were sampled and that all sixth graders in sampled schools were tested. The sampling and estimation procedure that provided the smallest variance was judged to be best.

To evaluate PPS sampling, it was assumed that schools were sampled with probabilities proportional to their sixth-grade enrollments (the data in the third column of Table A). To evaluate PPES sampling, a slightly more complex assumption was made. The measure of "size" used for a school was equal to the product of the school's sixth-grade enrollment and the average ability test score earned by the school's fifth graders (the data in columns three and four in Table A). While this product (sixth-grade enrollment times fifth-grade ability test score) might not have much meaning as an assessment statistic, it makes an excellent variable for PPES sampling since it is highly correlated with the total of sixth-grade reading achievement scores in a school.

The variances of estimators of average sixth-grade achievement in the district are given in Table B below:

Table B: Variances of Estimators of Average Achievement for Sixth-Grade Students in Anydistrict. Sample Size is 10 Schools from a Population of 21.

Sampling and Estimation Method	Estimator Variance
Simple random sampling of schools with unbiased estimation	21.790

Table B: (continued)

Sampling and Estimation Method	Estimator Variance
Simple random sampling of schools with ratio estimation	1.802
Sampling of schools with probabilities proportional to sixth-grade enrollments (PPS)	3.622
Sampling of schools with prcbabilities proportional to fifth-grade ability test scores (PPES)	1.358

From the data in Table B, it is clear that PPES sampling of schools is the most efficient of the four cluster sampling procedures. PPES sampling is slightly more efficient than simple random sampling of schools with ratio estimation, more than twice as efficient as PPS sampling of schools, and more than sixteen times as efficient as simple random sampling of schools with unbiased estimation. Efficiency is calculated from the ratio of estimator variances.

Although PPS sampling and PPES sampling are not consistent procedures, the variances of their estimators do decrease steadily as sample size is increased. Simple random sampling of clusters with unbiased estimation or with ratio estimation are consistent, so the variances of their estimators also become steadily smaller as sample size is increased. Thus, one can generalize from the data in Table B for all sample sizes that are substantially smaller than the population size. PPES sampling will be most efficient, simple random sampling of schools with ratio estimation will be next most efficient, PPS sampling will rank third in efficiency, and simple random sampling of schools with unbiased estimation will be very inefficient.

The formulas used to calculate estimator variances in this example can be found in many sampling texts, including Murthy (1967), Cochran (1963), and Hansen, Hurwitz and Madow (1953).

THE USE OF CORRELATES OF ACHIEVEMENT IN STATEWIDE ASSESSMENT

Paul B. Campbell

TABLE OF CONTENTS

It is generally recognized that to consider the results of student achievement measures without taking into account the conditions of learning frequently leads to inappropriate interpretation of the results and misguided action. A logical strategy to prevent these adverse effects is a systematic analysis of the conditions under which learning is attempted and the resources which are brought to bear on the learning attempt. In addition, direct consideration of condition variables is the first step in defining hypotheses about the causes of learning success or failure.

To accomplish this analysis, a two-stage model of assessment activity is recommended. In the first stage, data on both condition variables and student achievement should be systematically collected in such a manner that some competing explanations of the results are ruled out while others remain plausible. These data are statewide in origin, with comparisons available on specific conditions in contrast to specific organizations.

A careful analysis of the relationships which are found in the first stage is the basis for more intensive smaller scale studies. At this level, the unit of consideration moves from statewide data collection to an individual learning-unit study.

The methods for doing the large-scale data collecting and analyzing are illustrated by a number of studies and reviews that have been undertaken in recent years. Several of these studies have been selected because the variables they examined included those which share common variance with student achievement to an extent which suggests that fruitful causal hypotheses may be generated about the situations which these indicators or correlates describe. In none of these studies has the second stage been undertaken. Table 1 shows frequently occurring correlates which describe in part the variation of conditions under which learning is attempted.

TABLE I

Socioeconomic status variables

- Mother's occupation
- Father's occupation
- Mother's educational level
- Father's educational level
- Value of home
- Household income

Teacher variables

- Teacher's experience
- Teacher's salary
- Teacher's certification
- Student orientation in contrast to "subject" orientation
- Verbal facility
- Recency of training and level of education
- Job satisfaction - teacher turnover

School variables

- School site size
- Building age
- Percent substandard classrooms
- Library volumes per student
- Textbooks per student
- School size
- Student mobility
- Class size
- Number of special area teachers per student - lab facilities
- Average teacher time in guidance
- Length of school year
- Materials and supplies expenditures

Each of these correlates is significantly related to student achievement defined as some measure of verbal or mathematical performance, in one or more studies. To facilitate discussion they have been grouped in broad categories.

Socioeconomic Status Variables

The first group, socioeconomic status (SES) variables, shows a positive, strong relation to achievement in every one of the studies reviewed in which they were considered (Benson, 1965; Burkhead, 1967; Campbell, 1971; Coleman, 1966; Dunnell, 1971; Garon, 1971; and Kiesling, 1968). The methods of collecting such data vary from student questionnaires to estimates from census data. In many cases, there is strong reliance upon school records or school officials' opinions about the socioeconomic status of the neighborhood. The definition of the variable also ranges from family income through occupation to housing quality and parents' education. Regardless of the grossness of the measure, the positive relationship exists.

The important issue, however, is the interpretation of these findings. They do not establish that low or high SES is a cause of low or high student achievement. The SES variables are at best proxies for some set of experiences the student has had and through which he has developed his own unique coping style. More specific analysis of the factors associated with SES are illustrated in the work of Shipman (1971) on the mother-child interaction tasks. Her study suggests that language utilization patterns, which vary with SES, may be significant mediators of the learning experience. Another hypothesis is evoked by an unpublished study conducted by the author in 1970 of several very small high schools. Among these schools, the correlation between SES and achievement was nonsignificant. This study suggests that SES is not important where it does not have the effect of sorting the student body into social strata. In these schools, the small size of each grade seemed to limit the range of differential experiences of the students.

The data from SES studies in general suggest that qualitative differences in teacher-student interactions across the levels of SES are the most useful places to look for causes of variability in student achievement. These data also indicate, spanning as they appear to do the whole variety of educational experiences, that the causes of insufficient learning will not be easily found nor will solutions be quickly implemented. In pursuing the elusive causes of achievement variability, therefore, it is suggested that data on those forms of the SES variables which have the most direct relationship to the student's educational experiences, such as parents' education and allocation of community wealth to the educational enterprise, should be collected where possible.

Teacher Variables

The next group of variables which appear to relate to achievement are teacher related. They include training, experience, morale, salary, verbal facility, and attitude toward students. In general, although the relationship was much lower than the SES variables, teacher variables were reported significant in most of the studies (Benson, 1965; Burkhead, 1967; Campbell, 1971; Coleman, 1966; Goodman, 1959; Guthrie, 1971; Hanushek, 1968; James, 1963; and Kiesling, 1968). It is rare, however, for these variables to account for more than 10 percent of the student achievement variance. Three studies provide clues for possible causal hypotheses about teacher effects. Guthrie (1971) found verbal ability and job satisfaction to be significantly related to student achievement in a positive direction. Hanushek found a significant positive relation between the recency of teacher training in subject areas and the achievement variables. This training was not of the usual undergraduate type, but rather that acquired through facilities such as NDEA institutes. Kiesling (1968) likewise notes the negative effect of teacher turnover on student achievement. These data suggest that a teacher with up-to-date training in the subject matter he is teaching, who can communicate well with the students and is basically role satisfied, with best augment the educational experience of the students as measured by achievement tests.

School Facilities Variables

The final set of variables considered in this paper are those related to school facilities, broadly defined. They include physical characteristics such as building site size and building age. They also include arrangements which influence how teachers spend their time and characteristics which affect the school climate such as student independence or restrictions. This set of variables, like the teacher set, does not in general reach the strength of relationships found between student achievement and the SES variables. The results for school facilities variables are also less consistent from one study to another. Class size, for example, is sometimes positively and sometimes negatively related to student achievement. Of the thirteen studies reviewed, this variable was positive in four (Burkhead, 1967; Flanagan, 1962; Guthrie, 1959 and Shipman, 1971) negative in one (Dunnell, 1971) and did not achieve significance in the remaining eight. The variables of greater interest in this set are those which suggest a kind or quality of interaction between the student and his learning environment (including the teacher). A review of the commonly examined variables does not reveal any good candidates for this specification. Therefore, it is probably more profitable to relegate these variables to a secondary order to be considered only as they enhance or hinder the operations of the most important set, the learning variables.*

Process Variables

It is readily apparent that the correlates of achievement described in the proceding section are, at best, proxy or carrier variables which are not likely in themselves to be causative antecedents. It is also apparent that many such variables are not subject to alteration by the school. The

*The reader is urged to read the informative paper by Kiesling (1971) for a more detailed discussion of the condition variables and their analysis.

alternative for achievement improvement is therefore to be found in the process of education--those things which occur within the school's sphere of influence. This means the interaction among the teachers and students, with or without tangible materials as part of the setting, must be examined.

It is sometimes useful to classify processes according to function. Managerial or facilitative processes are those which bring about a setting in which an interaction can occur--e.g., reducing class size, building open classrooms, organizing modular scheduling, and providing elementary guidance personnel. They are a set of variables which frequently overlap the earlier defined school facilities but which may be more specifically directed toward programs which reflect the school's philosophy.

Learning processes, on the other hand, are those interactions which occur within the setting provided by the facilitative process and which involve the student directly. If, for example, the objective is learning to recognize the sense of a simple paragraph, a set of events must occur. The student must recognize most or all of the words. If he does not recognize all the words, he must be able to infer the meaning of the unknown from the known. He must select or infer the appropriate meaning of known words from the context, and, finally, he must understand the relationships among the words. It is probable that he does this by finding much that is familiar, enough new material to maintain his interest, and the thread of an idea that he wants to bring to closure. The teacher may interact in this learning situation by providing an "other person" model of interest in the idea. This role is best fulfilled by being genuinely enthusiastic, although a sincere interest in the learner may suffice. The teacher must also be sensitive to the ratio of the known to the unknown and must keep the unknown to a manageable level through the medium of providing the student with necessary information. In order to achieve this sensitivity, the teacher must be aware of the practices within the community which determine the meaning of certain behaviors, both verbal and physical (in the "body english" sense), and must be able to practice the necessary communication skills to convey and receive messages to and from the student. The acceptability and utility of such communication characteristics as level of voice (loud - soft) and choice of words (shut up - please be quiet) need to be understood. Finally, the teacher must provide reinforcement through reassurance on tentative but appropriate responses of the student.

Although much attention has been given to the teaching task, little is positively known about the nature of effective teacher-student interaction. It is here, to be functional, that assessment must make a contribution.

The documentation of teacher-student interaction and the analysis of its relation to student achievement is difficult, time consuming, and expensive. Although a number of observational techniques are available (see Flanders, 1966 and Medley, 1968), it is unlikely that such intensive observation procedures can be adapted to large-scale collections of data for statewide assessment purposes. However, statewide assessment offers a unique opportunity for examination of learning processes if a two-stage model is adopted.

In the first stage of this model, data are collected on student achievement and the condition variables of interest. The student achievement data are classified according to the levels or categories of the most explanatory correlates. In the second stage, a smaller sample of two types of classrooms within each classification, one markedly successful and the other markedly less so, are selected for intensive study.

TABLE II

Stage I

- Collect data on student achievement
- Collect data on condition variables
- Analyze the relationship of the two sets of data
- Classify achievement data by levels or categories of selected correlates

Stage II

- Select sample of classrooms from extremes within classifications
- Conduct intensive study of classroom interactions
- Collect data
  - Task card sort
  - Teacher questionnaire
  - Student questionnaire
  - Question formulating and alternative descriptions test
  - Teacher group interview

Under the assumption that effective teacher-student interaction may be mediated by the teacher's perception of the students as learners, the students' perception of the teachers as sources of information and support, the communication skills of the teacher, and the actual activities in which the group engages, five types of data collection are proposed. These are a card sort of classroom activities, a teacher questionnaire designed to assess empathy with students, a student questionnaire on perception of the teacher, a teacher test of ability to formulate questions and to explain concepts in a variety of ways, and a structured group interview directed toward sensitivity to student needs.

The card-sort device is used to provide a profile of the actual activities which go on in the classroom over time. Its development consists of three steps. First, a group of teachers is selected from the population of interest. For the purposes described [illegible] pulation would be teachers from each type of school which is [illegible] intensive study. These teac[illegible]s are contracted to provide a [illegible] activities in which they engage [illegible] random selection of days. The activities may range from teaching consonant blends to scolding the class for making too much noise. After this collection is complete, the activities are edited for overlap and clarity and printed on cards--one activity per card. The card sets are then reviewed for representativeness by another sample of teachers from similar schools, with the additions and deletions recommended by this group carefully considered. A preliminary analysis of activity differences between the types of schools may be conducted at this stage. This analysis can suggest possible interaction differences for further exploration. The principal data collection, however, secures from a new sample of teachers working in the intensive study schools a profile of activities they perceive to be occurring. On a random sample of days, these teachers sort the cards into two sets--those activities they did on the day in question and those they did not do. The cards are then tabulated by a clerk and retained until the next sample day arrives. As they occur, new activities can be recorded by the teacher on blank cards included each time in the deck. The relative frequency and the nature of the activities provides a picture of the common modes of operation in each type of school.

Such data must be supplemented by additional information. The teacher empathy questionnaire assists in this function by providing an assessment of the teacher's perception of her class. A series of statements covering a range of positive or negative attributes is presented. The teacher indicates which statements are most descriptive of the class. Examples of statements might look like this:

> "This class asks a lot of good questions."
>
> "This class has a lot of trouble learning, and and they just don't care."

Since teacher perceptions are likely to change as the class becomes more familiar, this scale should be administered a minimum of three times during the year to allow trends to appear.

A student scale provides a third component of the interaction situation. The teacher may be seen as a friendly adult to whom one can turn for assistance--in contrast to a task master who is to be avoided as much as the situation permits. A series of actions which a student may take involving the teacher are presented. The student indicates his likelihood of selecting each action in his present class. At the high school level, specific classes (e.g., English or chemistry) should be randomly assigned to the students enrolled so that each may react to a specific situation. The composite of all student responses will present a picture of the whole school. School data rather than individual data is desired; therefore, a tearoff tab indicating both the class and the student should be incorporated to protect the anonymity of both teachers and students.

A rational hypothesis concerning content-oriented teaching skills suggests that the ability to formulate appropriate questions and the ability to provide a variety of explanations of concepts are important factors. A test of these skills has been devised. The data it produces should provide additional documentation of the interaction scene which we believe produces learning. It is therefore included as a necessary component of learning process assessment.*

The final set of interaction data suggested for inclusion in the intensive study is derived from a set of structured group interviews. The school staff is assembled on several occasions and with several configurations of attendance. On each occasion the interviewer presents a topic for discussion, legitimizing, in turn, contrasting positions on the topic. Case studies or a series of film clips of classrooms in action are useful stimulators. The content should focus on the degree of understanding and acceptance among the participants.

The group configuration should include administrators on one occasion, teachers only on another, and a variation of teaching responsibility, if staff size permits, on still another occasion. The order of presentation should be rotated among the intensively studied schools to allow order effects to be assessed.

These data collection activities will provide a fix on the teachers' perceptions of activities actually occurring in their classes, their

*We are indebted to David Potter, Research Psychologist, Educational Testing Service, for some of the ideas presented here.

perception of the kinds of students they are working with, the students' perceptions of the kinds of persons their teachers are, and an assessment of the teacher's attitude toward the teaching task. From this set of data, the nature and quality of the student-teacher interaction may be inferred and either qualitatively or categorically described.

Finally, new achievement data is collected from the students of these schools. Variation in the student-teacher interaction data can then be compared with student achievement variation to discover interpretable relations. If the interaction components identified by the several methods of data collection are indeed those which influence the student's learning, several relationships should exist. Because, for example, SES varies with achievement, there should also be a joint variation of interaction components with SES and achievement. Otherwise, the experiences or prerequisites associated with SES which are influencing achievement have not been identified. If, on the other hand, interaction components are identified which vary with achievement but are independent of SES, then a genuine breakthrough may be at hand. Experimental verification is the next step. If the first alternative is true, however, the task becomes that of devising ways to alter the interactions in such a manner that they remain associated with achievement but become independent of SES. This, too, calls for experimentation. It is well to note that the interaction is the crucial factor, not the presence or absence of a certain process, such as style of presentation.

It should also be noted that there is at present very little documented difference between schools in terms of what they do. Therefore, it is quite likely that the modifications of teacher-student interaction patterns will have to be developed and introduced in order to bring about changes in achievement which are independent of the demographic and economic characteristics of schools.

In summary, the correlates of student achievement are useful in two ways. They describe conditions that vary in facilitating student achievement and help us to focus on areas in which it is fruitful to search for causes of learning difficulty, thus aiding the search.

One final problem remains. Statewide assessment is seen by many as a simplistic solution to the problems of improving quality without the time-consuming study proposed here. School reimbursement formulae, district comparisons, and legislative critiques are all part of the current picture.

The requirements of a good research design are not the only ones to be met. The political requirements may demand that some components that are unwarranted from a research point of view must be included as a necessary cost of conducting a meaningful study. The activities suggested here include several which will be deemed unnecessary by some and will be seen as a threat by others. The conditions for successful action must therefore be carefully established. The key principles to be followed in such endeavors are these:

> Involve affected groups early and significantly in the planning.
>
> Consistently reject blame placing and direct the available resources toward improvement.

iples are genuinely adhered to, the chances of a meaningful contribution to the quality of educational experience are good.

REFERENCES

Benson, Charles S., et al. State and local fiscal relationships in public education in California. Sacramento, Calif.: Senate of the State of California, 1965.

Burkhead, Jesse, Fox, Thomas G., & Holland, John W. Input and output in large city high schools. Syracuse, N.Y.: Syracuse Univer. Press, 1967.

Campbell, Paul B., et al. Educational quality assessment phase II findings: section 6: data analysis. Harrisburg, Pa.: Pennsylvania Department of Education, 1971.

Coleman, James S., et al. Equality of educational opportunity. Washington, D.C.: U.S. Government Printing Office, 1966.

Dunnell, John P. Input and output analysis of suburban elementary school districts. Research in Education, 1971, 6.

Flanagan, John, et al. A survey and follow-up of educational plans and directions in relation to aptitude patterns. Pittsburgh, Pa.: Pittsburgh Univer. Press, 1962.

Flanders, Ned A. Interaction analysis in the classroom. Ann Arbor, Mich.: Univer. of Michigan, 1966.

Garon, J.W. Personal and environmental factors related to the achievement of public secondary students in Washington Parish, Louisiana. Dissertation Abstracts International, 1971, 32, 2481.

Goodman, S. The assessment of school quality. Albany, N.Y.: New York State Department of Education, 1959.

Guthrie, James, et al. Schools and inequality. Cambridge, Mass.: Massachusetts Institute of Technology, 1971.

Hanushek, Eric A. The Education of Negroes and Whites. Unpublished doctoral dissertation, Massachusetts Institute of Technology, 1968.

James, H. T., Thomas, J. A. & Dyck, H. J. Wealth, expenditures and decision-making for education. Stanford, Calif.: Stanford Univ., 1963.

Kiesling, H. J. High school size and cost factors. Washington, D. C.: U. S. Department of Health, Education, and Welfare, Bureau of Research, 1968 (Project No. 6-1590).

Kiesling, H. J. Multivariate analysis of schools and educational policy Santa Monica, Calif.: The Rand Corporation, March 1971.

Medley, Donald M. Assessing the Learning Environment in the Classroom: A Manual for Users of OScAR 5-V. Research Memorandum 68-9. Princeton, N.J.: Educational Testing Service, 1968.

Mollenkopf, W. G., & Melville, S. D. A study of secondary school characteristics as related to test scores. Research Bulletin 56-6. Princeton, N.J.: Educational Testing Service, 1956.

Shipman, V. C. Disadvantaged children and their first school experiences: structure and development of cognitive competencies and styles prior to school entry. PR-71-19. (Prepared under Grant H-8256, Department of Health, Education,and Welfare.) Princeton. N.J.: Educational Testing Service, 1971.

DEVELOPING TESTS FOR ASSESSMENT PROGRAMS: ISSUES AND SUGGESTED PROCEDURES

John Fremer

TABLE OF CONTENTS

Overview

It would be difficult to overestimate the importance of selecting or developing tests, questionnaires, or other measurement instruments that will fulfill the goals of an assessment program. All a measurement instrument can do is permit the systematic collection of information. It is the job of planners and developers of assessment programs to insure that the information obtained is the kind of information that will be helpful in evaluating and making decisions about the status of education in a school district or state. Among the issues that need consideration are the following:

What should be measured?

What types of reports will be meeded?

Should newly developed or existing instruments be used?

How should new assessment instruments be developed?

This paper addresses each of these issues in turn and attempts to identify the major factors that will require attention and to offer possible design and development strategies.

What Should Be Measured?

The question of what to measure in an assessment program is one that has to be addressed both at a global and a specific level. Considering the global level first, one possible answer is that the program should assess the extent to which students, teachers, administrators, and other educational personnel, in short, the entire educational system is achieving the goals for--in short, system--education in a school district or a state. Most states and many school districts already have goal statements that have undergone a cycle of development and refinement. This process can be a very valuable one, particularly if parents and other members of business and community groups contribute to the task of setting and reviewing overall educational goals and establishing priorities. The participants in the goal-setting process are likely to become aware of the extraordinary breadth of goals that schools are being asked to address. These same participants might be able to serve as spokesmen for an assessment program that attempted to measure a wide variety of goals. Even in advance of a systematic review of goals for a school system or state, it is possible to make a fairly accurate prediction of the outcome of this review. A recent Cooperative Accountability Project report on State Goals for Elementary and Secondary Education (Zimmerman, 1972), for example, reveals considerable consistency among the goals statements of 35 states. Basic skills goals appear in many forms in the lists developed in the various states, just as they appear, without exception, in the goals of individual school systems.

The assessment program planner and developer has to go beyond the global level and has to deal with the problem of determining the priorities to be assigned to measuring the many goals for education. What emphasis should be given to the basic skills area, to other school subjects, to competencies that have particularly high survival value in our society, and to values and attitudes or other noncognitive attributes of students or of teachers? To what extent should the process of education--teaching styles, methods of classroom organization, etc.--be described and documented? These are difficult questions; moreover, they are not ones that should be answered wholly or even primarily by a technical assessment group. Many educators and members of the larger community have perspectives that will need to be brought to bear on the problems. It is clear, though, that the breadth of educational goals will require a sequential approach to assessment program development. The developer will have to start with some obviously important goal areas such as reading or communication skills or health or mathematics and concentrate his initial time and resources on adequate measurement of them. At the same time, long-term plans can be developed for addressing the other significant goal areas.

Assessment program developers have often initiated their programs with testing of reading and mathematical skills at one or more grade levels. Since these skills have high survival value and since a number of measurement instruments and approaches are available, this seems a quite reasonable way to start up a program. More difficulties can be expected if the program developer attempts to assess student or teacher attitudes and values; yet these noncognitive attributes are valued highly by educators, legislators, and private citizens. In order to create an assessment program that adequately reflects the goals for education in a school district or state, some measurement in noncognitive areas is recommended at the very beginning of the program. Awareness of measurement difficulties will encourage postponement of attention to the noncognitive areas. In this connection, it is worth considering the observation of Campbell, Bruno, and Schabacker (1972, p.3): "Although these noncognitive areas are admittedly more difficult to measure, in an assessment program they must not be ignored in the early phase, or they most likely will continue to be neglected as the program is enlarged."

When an assessment program developer is making plans for the initial assessment years, attention needs to be given also to the future direction of the program. It will often be difficult to predict the level of funding that will be available, but estimates will have to be made and long-term emphases identified. What goal areas can be added to the program in future years? What kinds of assessment cycles should be introduced? Should some goal areas be assessed yearly and others on an every-other year or every-third-year basis? Which tests can be reused in subsequent years, with or without some revision? Should some provision be made for workshops or special training materials for the users of assessment results? Questions such as these go beyond the initial question of what should be measured, but they set the stage for the issues addressed in the balance of this paper.

What Types of Reports Will Be Needed?

Once a developer has identified the areas to be assessed in the initial phases of a program, it is necessary to consider the reporting plans for the program. This job should be tackled as early as possible rather than left, as it often has been, until many other decisions about an assessment program have already been made. Decisions regarding the information to be collected and reported will directly affect instrument planning. Is it necessary, for example, to develop reports for individual students? If so, every student must sit for any tests for which such reporting is required. If, on the other hand, reporting will be done for groups of students, sampling procedures such as those outlined by Trismen (1972) and Jaeger (1973) can be employed.

Whether reporting is by group or individual, the nature of the reporting planned and the types of instruments needed to accomplish it need to be considered. The statewide and school district testing programs that are the forerunners of today's developing assessment programs report summary scores and sometimes subscores based on survey, norm-referenced achievement tests. The summary scores can be used to relate state or district results to national norms or to monitor the performance of groups over time. Such summary score reporting has received, however, a great deal of criticism on the grounds that it does not tell us what we need to know in order to take constructive educational action. A good deal of attention has been paid recently to the possibility of reporting assessment results for cognitive areas in terms of specific student competencies such as the abilities to:

-- address a business letter
-- pass a state driver examination
-- figure correct change, or
-- choose a nutritionally balanced meal

This same logic could also be used to call for reporting on attitudes or behaviors such as:

-- the number of nonrequired books of various types students read
-- the importance students attach to various rights expressed in Bill of Rights, or
-- The value students at specified grade levels place on certain environmental conditions

The calls for objectives-referenced, content-referenced or criterion-referenced tests have suggested that test developers need to determine precisely what students know or can do. Holders of this position indicate that critical objectives must be identified and associated measurement procedures developed along with judgmentally or empirically derived standards. These standards would permit a determination of whether or not students had achieved the objectives. Educators can then direct their efforts at those high-priority objectives that students have not attained. The argument has typically been framed in a way that calls for measurement procedures that yeild only "yes, he has" or "no, he has not" decisions regarding attainment of objectives. (Robert can or cannot identify the main idea in reading selections of a specified difficulty level.) The approach is easier to defend, however, if the concept of degrees of attainment of objectives is employed and if the probabilistic nature of measurement is kept in mind. (John can type 70 ± 10 words per minute.) Some advocates of objectives-referenced or criterion-referenced measurement have caused educational mischief by seeming to seek the unattainable goal of error-free measurement and thereby creating confusion regarding appropriate standards for measurement instruments, (adopting the untenable position that reliability and validity are concepts which are not applicable to criterion-referenced tests). There have been problems also with the setting of performance levels that will be taken as evidence that a student has attained an objective. Too often, arbitrary levels such as 85% or 95% correct have been used. Ideally, performance levels would be set with reference to some future situation such as the subsequent educational experiences that are planned for the student. Criterion-referenced testing would then indicate whether the student had achieved the skills and competencies necessary to perform well in the next program or unit of instruction.

The positive effects of the criterion/objectives-referenced testing movement, however, far outweigh the negative ones. One highly significant and positive outcome is that a comprehensive reevaluation of the purpose and uses of tests has been initiated. Developers of testing and assessment programs have had to consider carefully the types of information they can and should obtain from tests and to broaden their thinking about methods of reporting information to the various audiences for assessment results. For a discussion of reporting as it relates to criterion-referenced assessment programs, see "Developing a Criterion-Referenced Assessment Program" (Fremer, 1973).

Some assessment program developers have chosen to make use of the National Assessment pattern of reporting results on an exercise-by-exercise basis. This approach can be employed with any exercise or item, and it does seem to stimulate public interest. It is necessary, however, to contend with the problem of overinterpretation. It is natural for readers of such assessment reports to assume that the results from a single question provide insights that can be generalized to whole classes of skills and knowledge. Yet the results from another question tied to the same objective might well be dramatically different and thus lead to different conclusions. Careful pretesting of groups of similar questions can help. Items selected for reporting can be ones wit difficulties representative of the total group of items tied to an objective. Even when an item is chosen on this basis, however, the pool of available items may not represent adequately the pool of items that could have been written to measure the objective. It will always be necessary to recognize that measurement and interpretation involve errors and inferences that can lead to unwarranted conclusions. Qualified rather than absolute statements should be the goal of assessment program developers.

Reports of the proportions of students achieving specified educational objectives perhaps form a middle ground between total score reporting in terms of norms and the reporting of results on individual test items or exercises. (Reports for individual exercises for any given group can, of course, be related to the results for these exercises when administered to some norms group.) Reporting of the proportion of students achieving specified objectives can be the outcome of the administration of homogeneous sets of items or exercises aimed at these objectives. This work or task sample approach has been the typical route to reporting by objectives. It is also possible, though, to use available survey achievement tests to make estimates regarding the proportions of groups of students that have attained specific objectives. The results of survey achievement tests need to be related by experimental procedures to the behaviors or competencies that are of interest. Such an approach involves the use of information on a number of aspects of subject-matter mastery to estimate mastery of particular skills. This idea is developed in a report entitled "Criterion-Referenced Interpretations of Survey Achievement Tests" (Fremer, 1972).

Should Newly Developed or Existing Instruments Be Used?

A developer that has selected assessment areas and decided to use particular types of instruments and reporting approaches will almost certainly have made these decisions with some reference to his knowledge of available instruments and his estimate of the feasibility of developing new instruments. Regardless of the areas chosen, there are likely to be some instruments that would have a claim to appropriateness on the basis of their titles or descriptions appearing in journals or publishers' catalogs. In the area of reading, for example, the Test Collection at Educational Testing Service had collected some 700 tests as of November, 1973 from all parts of the country and the world. Whatever grade level was planned for the testing of reading skills, a stack of tests of mixed origins and quality could be identified. Knapp (1972, 1973) has provided an indication of the availability of instruments in the noncognitive areas of school-based attitudes and self-concept. Other sources provide listings and evaluations of existing tests. The following are some helpful sources:

Mental Measurements Yearbook Series (Gryphon Press, Highland Park, New Jersey)

The volumes in this series include descriptions of tests critical reviews, publishers' directories, and bibliographical references.

1. Mental Measurements Yearbooks (MMY)
2. Tests in Print
3. Reading Tests and Reviews
4. Personality Tests and Reviews

CSE: Elementary School Test Evaluations and CSE-ECRC Preschool/ Kindergarten Test Evaluation

These volumes include ratings of tests on a number of criteria. They are published by the Center for the Study of Evaluation and the Early Childhood Research Center, UCLA Graduate School of Education, Los Angeles, California.

NCME Measurement News

This newsletter of the National Council on Measurement in Education contains general articles on testing issues as well as announcements of new tests and lists of test reviews.

Test Collection Bulletin (TCB) (ETS, Princeton, New Jersey)

This is a quarterly digest of information on tests and services which generally have become available after the publication of the most recent Mental Measurement Yearbook. It describes both commercially available tests and tests used experimentally. The Bulletin does not evaluate the tests listed, but it does provide references to test reviews.

The abundance of existing tests places a burden on the developer in that attention needs to be paid to their evaluation. In this connection, a committee of reviewers representing the groups who contributed to the goal-setting process can be helpful. It is likely to be the case that no existing instrument would be ideally appropriate for any given assessment program; yet, the best available instrument may be judged acceptable, particularly if time, staff, and budgeting constraints permit no other alternative. The use of nationally standardized tests may still be appealing even when the schedule and budget would permit local development efforts. The fact that standardized tests have had extensive editorial and subject-matter reviews and careful pretesting can be of value in defending a program.

The items in such tests could be matched to educational objectives and reporting carried out for appropriate clusters of test items. It should be recognized that monies not used for development in one assessment area can be allocated to other areas. Use of a standardized reading or mathematics test could therefore free up funds for work in attitudinal or other noncognitive areas. Ideas for new approaches to testing in either cognitive or noncognitive areas could be explored and perhaps carried to the point of pretesting. It would also be possible, for example, to supplement an existing standardized test with newly developed materials covering aspects of content not emphasized in the best available standardized tests. A set of questions on aspects of arithmetic important to twelfth graders as prospective consumers, renters or buyers, and income tax payers could be added to a conventional survey mathematics test. Questions on local or state history or government can supplement the content of a more global Social Studies test.

Whatever the balance of existing or newly developed materials included in an assessment program, it will be desirable to provide some time for pretesting of new material. It is often necessary to fight for blocks of testing time, and teachers and administrators are understandably reluctant to add to the minimum that was granted during the first year of a program. Failure to seek enough time for the tryout of materials, though, can remove a convenient mechanism for gradual evolution of a program.

How Should New Assessment Instruments Be Developed?

The instruments used in assessment programs and their methods of development are likely to receive a great deal of critical attention from educators, school board members, legislators, private citizens, and the press. It is important, therefore, for program developers to adhere to high measurement standards in the design and implementation of the assessment program. This goal will most likely be achieved if staff can be identified and utilized who have both extensive training in measurement and statistics, and first-hand experience with the development of testing and assessment programs. A school district or state assessment team can include some relative newcomers to the field of assessment, but it must have a solid core of old hands.

Any project is likely to succeed or fail on the basis of the quality of the staff who are running it, yet even a good team is not sufficient. The odds that a good staff will do a good job will be heavily influenced by the extent to which adequate planning takes place. This paper identifies general areas of assessment program development that will require careful thought. Each assessment situation will present its own special problems, but to ignore any of the general issues listed is, in the judgment of this writer, to court trouble. The points to be considered are grouped into the following six areas:

Initial planning and allocation of responsibility
Development of instrument specifications
Item development
Pretesting
Use of item analysis
Final test assembly

Each of these areas is considered in turn.

Initial planning and allocation of responsibility:

1. Identification of all components of the instrument development project -- This step involves the participation of all project staff, supplemented by external consultants with skills that round out talents of the project team. An extremely useful source of information in this connection is the chapter "Planning the Objective Test" by Sherman Tinkelman in Educational Measurement, edited by Robert L. Thorndike.

2. Development of a schedule for the completion of the steps -- This task is most readily accomplished by working backward from identified administration and reporting dates. The length of time needed to accomplish each step is determined using whatever sources are available. The identification of critical sequences can often be facilitated through the use of PERT charts (Wagner, 1973) or other diagnostic or tabular methods of presenting data.

3. Fixing clear lines of responsibility -- The overall Project Director will assign responsibility for aspects of the work to his staff on the basis of their experience and competencies. It will be valuable to not only establish clear lines of final responsibility, but to provide a second or back-up person for every task. The back-up person would review the primary person's work and remain sufficiently involved so that he could step in temporarily in the event of staff changes, illnesses or the like. The use of the Project Director as the only back-up person is to be avoided wherever the size of the staff exceeds perhaps five people. A written statement of responsibilities will be useful for large working groups. Such a statement can help other departments or agencies work efficiently with the project team.

4. Relationship to long-term goals -- Long-term goals usually receive a good deal of attention in the course of making initial program decisions, such as the identification of goal areas for early assessment. It is difficult, however, to continue to keep the long-term goals in mind when making the many specific decisions that design and implementation of a program require. Members of the project team can try

to raise questions of long-term impact when they review their own work and that of their colleagues. An advisory group can also be helpful, particularly if an evaluation of the relationship of present plans to future goals is made part of their charge.

5. Possibilities for multiple uses of assessment program data -- Most assessment programs are developed with more than one use for the data in mind. Emphases do vary, and one program will be focusing primary attention on the provision of global information to administrators, whereas another program will be devoting primary attention to the evaluation of particular programs. It will often be possible to serve an overall major goal quite effectively and still make provision for additional uses of the resultant data. The two examples of global evaluation at the school district or state level and individual program evaluation, for example, are quite compatible. It is true that the program evaluator will need to compare the content of an instrument used in the assessment to the objectives of the particular program, but the assessment program developer can help by providing ready access to the considerations influencing the instrument development process. The local evaluator can be further assisted if the instrument administration pattern produces individual scores that can be aggregated in various ways at the local level.

6. External control of aspects of a program -- The Program Director of an assessment program will want, generally, to maintain the level of control permitted by his position in an administrative system. Consideration should be given, however, to delegating to an external group such as an advisory committee responsibility for certain components of the assessment. In the development of instrument specifications, for example, a committee of educators might be given a decision-making rather than advisory function within limits defined by the Project Director and his staff.

Development of instrument specifications:

1. Involvement of many groups -- The specifications for assessment instruments should probably never be developed solely or even primarily by an internal staff group. Even when a school district or state has a large assessment staff with many talents and perspectives, the results of its unaided efforts will be judged unacceptable by the significant groups who were not represented in the specifications-development process.

2. Early and continuing external involvement -- The later one waits to involve an external group in the assessment process, the more likely it is that the group will resent the possible implication that they are being called in to "rubber stamp" the plans of internal staff. It is difficult to make changes late in the process of instrument development without bypassing desirable review and quality control steps, so the program developer is likely to resist suggestions for change. An advisory group that is involved early in the development process will have the ability to help formulate those aspects of specifications that are easy to identify and to reach agreement about, as well as the ones that result in disagreement and can only be handled through compromise. An advisory group that has worked through this process will be more likely to defend than to criticize the resultant specifications.

3. Covering all types of specifications -- Discussions of test specifications often center narrowly on subject-matter content for cognitive tests. When considering attitudinal or other noncognitive areas, it is necessary to expand the concept of "content" objectives to cover the classification of behaviors and occasions. For both cognitive and attitudinal instruments, it is also essential to go beyond content specifications to consider such additional categorizers as the following:

   a. Statistical Specifications -- Appropriate statistical specifications or selection criteria for individual items and for sets of items need to be developed. Item difficulty will be significant if a norm-referenced instrument is being constructed as this statistic will help guide the development of a test that will differentiate among the levels of skill represented in a particular population. Item difficulty information will also be valuable if an objectives-referenced or criterion-referenced test is being developed. In this latter situation, item difficulty can serve as a check on the reasonableness of particular objectives for various grade levels. It will also be useful to assess the degree of agreement among the difficulty levels of items judged to be equivalent measures of the same objective. If the items fail to yield results congruent with expectations, the items may be testing different attributes than those intended. Item to total-test or to subscore correlations will be useful as an index of item homogeneity and as a stepping stone to the evaluation of score reliabilities. Since reliability indices permit an estimate of the likelihood that a similar score would be earned on a parallel set of items, this information is essential to an adequate evaluation of any test. It has been suggested that items for criterion-referenced tests should be selected from among those items that are sensitive to instruction (Roudabush, 1973). Even in this situation, though, scores would have to be stable or reliable in the absence of instruction for the results of testing to be meaningful.

b. Question-Type Specifications -- A number of practical constraints have led assessment program developers to rely primarily on paper and pencil, machine-scorable question types. Each assessment program developer needs to consider, though, the possibility that other approaches would be more appropriate to the goal area under consideration. Consider, for example, the measurement of writing ability. Objectively scorable item types have been developed and validated against actual writing ability; yet it is clear that writing ability can only be measured directly through exercises requiring writing. Inclusion of actual writing exercises creates a need for professional scoring of exercises, but the likely increase in the credibility of assessment results may well justify the expense. If the writing exercises are administered only to a sample of students, the expense of scoring need not be very great.

c. Stimulus Material Specifications -- In addition to reviewing the possibility that a variety of question formats might be feasible, attention should also be given to the use of other than written stimulus material for questions. Tapes, films, and slides might be employed with samples of students or with an entire assessment population. Test administrators can be trained to read certain materials, speak certain sounds, or make use of apparatus of various types. Clearly, budgetary factors must be taken into account, but an assessment program must provide you with needed information if it is to be of value. Some types of needed information cannot be obtained by the least expensive testing formats.

d. Cultural Values Specifications -- Cultural values are usually thought of as the province of some special area of testing such as citizenship if they are thought of at all. Yet tests do communicate values to students, and it is well to consider this fact when designing the test. What provision is going to be made to represent various groups in the test-development process? What guidelines for test content will keep attention focused on an appropriate balance of contributions from many different facets of subject-matter fields? What values will be implied by the stimuli and questions?

e. Other Specifications -- The foregoing "special" categorizers do not exhaust the list of item and total-test attributes that an assessment program developer needs to be sensitive to. They may be helpful, however, as indications of the breadth of concern that is essential to successful program development. Each program developer will need to work with fellow staff members and with outside people to identify the additional areas for which specifications will be needed.

Item Development:

1. Specifications first — Item development is such a difficult, important and time-consuming part of assessment program development that there is a strong tendency to want to begin item writing without adequate attention to detailed program specifications. It is essential, though, to design content specifications that clearly identify what is to be measured <u>before</u> item development commences. In some instances, this may require the elaboration of detailed educational objectives that are implied by or subsumed under existing educational goals for a state. Such work on objectives is an essential part of assessment program development when results are to be reported on an objective-by-objective basis. It is a possible but not mandatory procedure when more global reporting is intended.

2. Use of existing models — Item development for assessment programs is often initiated because of dissatisfaction with existing tests and the items contained therein. It is inefficient, nevertheless, to ignore existing tests as a source of models, or at least ideas, for new items or exercises. Much can be gained by collecting existing tests and taking a hard look at what is or is not desirable about the constituent items. One can then employ in a new test any format or approach that seems suitable and identify the undesirable features that the new items will be sure to avoid. It is possible to evaluate the extent to which new items are actually better than the existing ones. The new items can be mixed with the "bad" items from existing tests. All tests should be typed on cards or standard forms so there is no clue to origin. Reviewers can then be asked to rank or assign a quality rating to every item. If the new items are indeed better, they should receive more positive evaluations. (This tactic is not recommended for assessment program developers with fragile egos.)

3. Staff for item development -- Every assessment program developer will need to decide who should write and review the needed test items. Are there staff available in the school district or state Department of Education? Can a local school district obtain help for its program from the state Department of Education? Conversely, can the statewide program draw on school districts or colleges for help? What part should "outside experts" from test publishers, research laboratories or centers play in the process? These are questions that each assessment program developer must answer in the context of the options available to him or her. Whatever the direction taken, though, staff experienced in instrument development must play a major role in the item development process. It is true that there is room for individuals with all levels of prior experience in an item-development group, including some staff who are receiving their first on-the-job training in the area. There have to be experienced hands on board, however, if training is to be successful.

4. Training item writers — How can one go about training item writers? One effective technique is the item-writing workshop. A good workshop provides participants with training in the elements of successful item writing and incorporates a goodly amount of actual writing and reviewing experience. Generally, two or more days will be required so that two or three full cycles of item writing, review, and revision will be possible. The actual writing of items is almost always best accomplished by having item writers work independently. Item review and revision, on the other hand, should involve both individual and group work. The group sessions are opportunities to discuss different aspects of items and to explore alternative approaches. The individual sessions permit the most efficient production of materials.

It is useful if participants can be prepared in advance for productive learning sessions through the use of background materials that clearly define the item-writing task and permit prior familiarization with both terminology and the fundamentals of technique. If possible, trainees should even write some items to the appropriate specifications and bring them to the first training sessions.

One component of an item-writing session should be the attempt to generate ideas for items or exercises that can be further developed at a later time into usable items. This is a strategy that has been employed in the development of exercises for the National Assessment of Educational Progress. Some review of the exercise-development procedures employed by the National Assessment is recommended for anyone considering individual exercise reporting. (See Finley and Berdie, 1970.)

5. Types of reviews needed — A full treatment of the process of item review needs to touch on many different purposes for reviews. Three such purposes are identified below:

   a. It is obvious that assessment items need to be appropriate measures of the objectives of interest. To meet this deceptively simple criterion may require considerable statistical work, but it also calls for reviews by individuals thoroughly familiar with the objectives and with the subject-matter domain of interest. Such reviewers can certify the appropriateness and accuracy of items and with guidance from measurement-trained staff can help evaluate the scorability and reportability of exercises that require judgment, (any nonobjectively scorable exercise).

   b. Despite all the contributions that subject-matter specialists can make to item review, yet another review is needed for consistency of style and clarity of expression. This review is best entrusted to a skilled specialist who has the same role for all assessment instruments and items. This procedure facilitates uniformity of format and style. If possible, this same editor/reviewer should also hold responsibility for controlling the readability level of items and of associated directions and explanatory materials. Only if students can understand the tasks posed to them, is it reasonable to view item and test performance as a reflection of their developed competencies.

   c. As a final suggestion in this area, developers should consider a possible review role for parents and other concerned citizens—representatives of a crucial audience for assessment programs. If parents are to contribute effectively they should be brought in early in the review process and should be given appropriate background information about program purposes and procedures. Nontechnical reviews of this nature can serve a valuable public relations function and can bring helpful information on issues such as the importance ascribed to various types of potential assessment material and the offensiveness and controversiality of exercises. Some input on these issues can also be obtained by giving students a role as reviewers. Students can be given an opportunity to comment on items as part of a post-pretesting session, when pretesting is included as a step in item and instrument development.

Pretesting:

1. Use some type of pretesting -- Some form of pretesting is a very desirable, perhaps even essential, component of an effective assessment program. As is noted in later sections of this paper, pretesting provides valuable information to the program development staff, but it has other benefits as well. Given the careful public scrutiny that can be brought to bear upon assessment program instruments, the protection against faulty items afforded by pretesting is very welcome. State assessment programs are often legislatively mandated with relatively inflexible time schedules, so the first year of a program may have to proceed at a pace that precludes certain types of pretesting. Even in these circumstances, however, some form of item tryout along the lines of those described in this paper is almost always possible. The problem is one of determining what kind of pretesting is possible within time and budget constraints, constraints which will apply also to the school-district assessment program. In the limiting case wherein almost no pretesting is possible, two strategies ought to be considered. The first involves the use of items for the initial assessment battery that have already been pretested on a population similar enough to the assessment group so that judgments of item

qualities can be made with some confidence. The second strategy is simply that of treating the first year of the program as a pretest, even if scores have to be reported, in that information gained therein can be used to revise the instruments for use in subsequent years.

2. Developmental trials -- One form of pretesting that requires very little in the way of time and money is item tryouts conducted by the original item writer with a small number of students. In this situation, the items can be administered on an individual basis, and the students can be interviewed by the item writer. This type of pretesting does not lead, typically, to the development of item statistics. Rather, it permits an opportunity for the clarity of wording of questions and directions to be checked by that individual who is most familiar with the intention of the item. The item writer can observe, to the best of his or her ability to do so, what it is that the students seem to be doing when they answer or solve the problem or question. Do they appear to be carrying out the process originally intended by the item writer, or is there some other method of obtaining the answer that is at variance with the objective for which the item is intended? An item, for example, that is designed to require a student to use insight or to synthesize data from many sources, would be judged suspect if students seem to be answering the items solely from factual recall.

   Developmental trials provide an excellent opportunity to discover vocabulary or phrasing of questions that is simply too difficult for the age level for which the items are intended. In order to achieve the maximum benefits from developmental trials, it will be desirable to conduct them with students comprising the lower end of the competency range at the age or grade level under consideration. Another potential benefit from developmental trials is an opportunity to obtain a first fix on the amount of time students will need to respond thoughtfully to the test items.

3. Small group trial -- Perhaps the next level of pretesting in terms of time and money required after developmental trials is the "quick and dirty" administration of items and test materials to small groups

of students without carrying out the same level of quality of production of test materials as is intended for the final administration. The use of spirit masters or xerography may well be economical here if the number of cases involved is sufficiently small. The small group trials could be limited to an examination of the effectiveness of directions for items in communicating the nature of the task. At a slightly larger level of involvement, the pretesting could incorporate sample items from each of the various types of items planned for inclusion in the final instruments. The items chosen for pretesting should be representative of other items in the domain, including some at the upper limits of complexity and difficulty.

4. Full-scale tryouts -- When it is possible to produce test material at about the same level of quality as the final instrument and to try out these items with groups clearly representative of the actual population, it will be very much to the advantage of the test developer to do so. There will always be an interest, of course, in holding down costs, so consideration should be given to methods of pretesting that are efficient. One opportunity to be explored in this connection is the use of a pretest or experimental section that can be added on to the regular battery in an existing testing or assessment program. As was noted earlier, inclusion of a pretest section in the first assessment battery is likely to be easier than trying to add one in subsequent years. When this opportunity exists, the costs of locating an appropriate sample and of setting up the administration conditions can be eliminated. The additional costs for pretesting may still be substantial as it is necessary to develop the items and to arrange for the production of the materials to be included in the experimental section.

5. Trend-line pretesting -- For some purposes pretesting may be most useful if it can be conducted on more than one occasion. When attitude measures are being developed, it is often useful to attempt to trace the development of attitudes over the course of the particular age or grade that will be the subject of study. It is often found that at early ages student attitudes simply lack the stability that would make successful attitude measurement possible. It is better

to find this information through pretesting than to incorporate it into the final assessment program only to have to explain away a failure to report results. The use of pre- and post-instruction pretesting can also be explored in the cognitive domain. Part of an assessment program might then be focused on those cognitive areas that are known to be sensitive to the types of instruction now being employed in most of the schools in a state or school district. This technique can be employed either for a reporting-by-objectives assessment program or for a global reporting program. For information on the kind of item analytic procedures that might be used in developing items that are sensitive to instruction, see Roudabush (1973). The issue of appropriate item analyses is treated in more detail in the next section of this paper.

Use of Item Analysis:

1. Item difficulty -- As was noted earlier, item difficulty information is valuable for both norm-referenced and criterion-referenced test development. Whenever items can be scored right or wrong--as is the case with most multiple-choice items in the cognitive domain--item difficulty can be determined. Similarly, one can determine the difficulty of sets of items, such as all the items related to a single objective or all the items relating to a single domain (in other words, a total test score). This kind of information is a necessary prerequisite for assessment program developers who need to build equivalent test forms for use in subsequent years of a program.

2. Item correlations -- One of the most useful statistics for evaluating the adequacy of test items is the item-test correlation. This index can indicate to the developer the extent to which any individual item is measuring about the same thing as other items in a cluster or in the total test. The typical values associated with item to total-test correlations will vary as a function of the homogeneity of the content covered by the test as well as with the heterogeneity of the group sitting for the test. The developer will have to become familiar with the range of correlations to expect for any given subject-matter domain or attitudinal area. One immediate use of the item to total-test correlation is to identify those items that require careful editorial

examination for possible ambiguities and technical inadequacies. Whenever an item is included with other items in an item cluster or in a total test because it appears on logical grounds to be a member of the same subject-matter domain or noncognitive attribute, a very low positive item to total-test correlation or a negative item to total-test correlation is an indication that the item is measuring something other than that intended by the developer. Most frequently, the developer will discover that the item is being interpreted in a manner not originally expected or that there is some irrelevant characteristic which is preventing the item from functioning as intended. It will often be possible to revise such an item and use it after a re-pretesting confirms that the problem has been corrected. It will sometimes be the case, however, that an item will prove unrelated to other items for reasons that are not at all apparent even after an intensive study of the content of the item. The inclusion of the item in a test where it will merely be contributing to some total score is to be discouraged. Results for the item, though, may suggest hypotheses about student competencies that can be followed up in experimental studies.

In certain circumstances, the results of an analysis of item correlation may suggest that some subset of items should be treated differently from the remaining items. This outcome is highly likely when item performance is correlated not only with total-test score but also with other items that are thought to be measures of the same objective. This procedure can make it possible to assess the homogeneity of items thought to measure the same objectives. If an item is no more highly related to its own cluster than to all items taken together, there is little evidence for thinking that that objective is indeed measured uniquely by the items that seem on logical grounds to be closely related to it. Further evidence for objectives' interrelatedness can, of course, come from the procedure of correlating item-cluster scores with other item-cluster scores. If sufficient funds are available, factor-analytic procedures can also be employed to refine the clusters of items related to individual objectives.

3. Analysis of options -- Test developers will be greatly aided if they employ item-analysis programs that indicate the number and relative test performance level of the students choosing each option to a multiple-choice question. This method of analysis permits the ready determination of those answer choices that are acting to depress item to total-test correlations, and can often suggest the nature of the ambiguity or misinterpretation that is interfering with the functioning of items. Such analyses may also suggest other questions that would be more appropriate measures of a given objective and can shed some light on the nature of student misconceptions or problems of interpretation.

4. Development of scales -- The analytic techniques already mentioned can be combined in order to develop knowledge tests with clusters of items related to somewhat independent objectives. They can also be used in the attitudinal area to sharpen measurement of given attitudes, interests, or values. It is, of course, inadvisable to rely solely on statistical data to refine reportable scales in these noncognitive areas, but statistical data can suggest hypotheses regarding the organization of a student's beliefs and positions, which will permit a sharpening of potential scale definitions. A scale defined in this manner, however, will require careful scrutiny to insure that the final collection of items to be reported in terms of a single score do indeed bear a close relationship to each other that is consistent with the developers' understanding of the nature of the attribute being measured. What the assessment program developer has to avoid is a kind of blind empiricism which could lead to the reporting of scores that have <u>no</u> theoretical organization but which "hang together" in only a statistical sense. It ought always to be possible for the developer to state clearly what a high score on any collection of items should mean and what a low score on that same collection of items should mean.

5. Triangulation -- One invaluable aid to the development of scales in the attitudinal domain and to sharpening one's definition of content areas in the knowledge domain is the collection of independent bits of information regarding the same competency or attribute. This procedure, which has been called "triangulation" by some writers,

can involve using more than one type of item to measure an attribute. It can also make use of nontest indicators such as teacher judgments or counts of observable behaviors as one line in the triangulation procedure. Consider, for example, the possible assessment area "attitude toward reading." Two different types of items, one requiring direct statements from students and the other requiring responses to objective questions, might be employed. In addition, teachers might be asked to judge how positive their students were toward reading, and the school library might be asked to maintain records of the extent to which these same students borrowed and read books. If the information contained from these three sources tended to yield similar conclusions regarding individual students, one could be fairly comfortable that attitude toward reading rather than some other attribute had indeed been measured.

Final Test Assembly:

1. Components of test assembly -- Final test assembly encompasses activities such as the review, selection, revising, editing, formatting, and organizing of the items or exercises for the instruments of the operational assessment battery.

2. Clarifying final responsibility -- One individual should have primary responsibility for each instrument in an assessment battery regardless of the number of people contributing to the process and whether or not an outside group has contracted for the task. The involvement of a continuing committee working with the assessment program staff is recommended at the time of final test assembly as it is at this point that all earlier work is synthesized. The final test assembler, though, needs to have the authority to make the many decisions which will come up as the test nears completion.

3. Final item review -- What precisely are the tasks facing the responsible individual, his cooperating staff members, and the committee? One significant task is final item review. All information available about the items in the pool should be collected in a convenient form and each item reviewed in the

light of this information. One useful strategy in this respect is the preparation of spiral notebooks with items on one page and the following on facing pages:

- objective or area of specifications covered
- pretest information (if any)
- previous reviewers' comments
- correct response or scoring guide

Whenever items are to be reviewed, it will be useful to keep the correct response separate from the text of the item so that the reviewer can choose or formulate an answer and tben check it against the official key or scoring guide. If the individual with primary responsibility for a test concludes that an item or exercise is ambiguous or that it lacks a single correct key, that item should not be used in a test, irrespective of the quality of its pretest statistics. Similarly, an item in the attitudinal area that appears to be subject to irrelevant interpretations should not be used as part of a scale, again regardless of its pretest statistics.

4. Meeting assembly specifications -- The process of screening out items because they are judged to be inadequate by reviewers can have the effect of reducing the pool of items in some areas so that it appears impossible to meet the original specifications for a test. At this point, it is necessary to consider whether some previously rejected items can be revised, whether additional materials can be created or whether the intended scope of the test will have to be reduced. If it proves necessary to narrow the focus of a test, it will be important to describe just what is and is not being measured by the instrument that is used operationally. In rare cases, the screening and culling process may produce a significantly larger body of items than is needed to meet specifications. In such instances, one can sample from the pool in such a way as to leave a set of items that is approximately equivalent to the items used, thus creating the possibility of a parallel form for subsequent use. When a test is designed to show the large variations in competence that are likely to be present in populations, statistical considerations will often help the developer determine which items

to use. When both statistical and content dimensions need to be satisfied, few developers will find that there is a large surplus of items in many areas.

5. Coordination with test-production staff -- When organizing the final test-production set of items into total tests or sets of related exercises, it will be useful to consult with the staff members who will be responsible for producing many copies of the final test. Decisions regarding the layout of items on pages and the order and sequence of items may have considerable implications for the total cost of producing the final package. No assessment program developer will be comfortable with page layouts that introduce complexities to questions beyond those necessitated by the nature of the task. The use of type too small to permit easy reading or excessive packing of questions into pages may undercut the most careful effort to produce quality instruments. Even when consultation with production staff is possible prior to final page layouts, the individual with primary responsibility for an instrument should review the printing masters prior to the test production runs. It is at this point that one is likely to discover such horrendous outcomes as the fact that stimulus and response materials have been inadvertently separated, the options for multiple-choice questions are improperly sequenced, or that no space was provided for students to respond to free-response questions.

6. Documentation -- Although the task of test assembly is often so complex and demanding that it is difficult to set aside the time to keep accurate records of decisions made, the absence of such records can often create substantial problems for the program developer. In general, every effort should be made to pick up potential errors as early as possible in the development process so that last-minute changes that will not receive a significant number of later reviews can be avoided. There will always be a need, however, for changes to correct errors that are discovered at the eleventh hour. Careful documentation of the reason for the change, the nature of the change, and the steps that were taken to inform all significant people will reduce the probability of catastrophic errors. Imagine the consequences of rearranging the questions for a test at the last minute,

so that the numbers of different items were changed, without notifying the individual who has already prepared the official scoring key for the test.

Final Comment

This paper has attempted to provide practical guidance to those individuals responsible for selecting or developing instruments for assessment programs. The suggestions that have been offered are all based on first-hand experience with the task of developing such instruments; yet it is clear that in any individual situation other possible courses of action could have been suggested and, if followed, might have yielded quite satisfactory results. There is no one correct way to develop an assessment program, but the enterprise has so many facets that specific suggestions regarding ways that a number of component tasks could be handled may be of value.

REFERENCES

Buros, Oscar K. (Ed.) The seventh mental measurement yearbook. Highland Park, N.J.: The Gryphon Press, 1972.

Buros, Oscar K. (Ed.) Tests in print: a comprehensive bibliography of tests for use in education, psychology, and industry. Highland Park, N.J.: The Gryphon Press, 1961.

Buros, Oscar K. (Ed.) Reading tests and reviews. Highland Park, N.J.: The Gryphon Press, 1968.

Buros, Oscar K. (Ed.) Personality tests and reviews. Highland Park, N.J.: The Gryphon Press, 1970.

Campbell, Paul B., Bruno, Nancy L., and Schabacker, William H. Statewide assessment: methods and concerns. Princeton, N.J.: Educational Testing Service, 1972.

Finley, Carmen J., & Berdie, Frances S. The National Assessment approach to exercise development. Ann Arbor, Mich.: National Assessment of Educational Progress, 1970.

Fremer, J. Criterion-referenced interpretations of survey achievement tests. Test Development Memorandum 72-1. Princeton, N.J.: Educational Testing Service, 1972.

Fremer, J. Developing a criterion-referenced assessment program. Paper presented at annual meeting of the National Council on Measurement in Education, New Orleans, February, 1973.

Hoepfner, R. (Ed.) CSE elementary school test evaluations. Los Angeles, Calif.: Center for the Study of Evaluation, UCLA Graduate School of Education, 1970.

Hoepfner, Ralph, Stern, Carolyn, and Nummedal, Susan G. (Eds.) CSE-ECRC preschool/kindergarten test evaluations. Los Angeles, Calif.: UCLA Graduate School of Education, School Evaluation Project, Center for the Study of Evaluation, and Early Childhood Research Center, 1971.

Jaeger, Richard M. A primer on sampling for statewide assessment. Princeton, N.J.: Educational Testing Service, 1973.

Knapp, J. An omnibus of measures related to school-based attitudes. Princeton, N.J.: Educational Testing Service, 1972.

Knapp, J. A selection of self-concept measures. Princeton, N.J.: Educational Testing Service, 1973.

Roudabush, Glenn E. Item selection for criterion-referenced tests. Paper presented at annual meeting of American Educational Research Association, New Orleans, February, 1973.

Tinkelman, Sherman N. Planning the objective Test. In R. L. Thorndike (Ed.), Educational Measurement. Washington, D.C.: American Council on Education, 1971. Pp. 46-80.

Trismen, Donald A. Sampling techniques for assessment. Princeton, N.J.: Educational Testing Service, 1972.

Wagner, Andrew R. What you always felt you should know about PERT, but were afraid to find out. Research Memorandum 73-15. Princeton, N.J.: Educational Testing Service, 1973.

Zimmerman, A. Education in focus: a collection of state goals for public elementary and secondary education. Denver, Col.: Cooperative Accountability Project, 1972.

STATEWIDE ASSESSMENT:
METHODS AND CONCERNS

Nancy L. Bruno
Paul B. Campbell
William H. Schabacker

TABLE OF CONTENTS

APPENDIXES

Introduction

Traditionally, the method of assessing the quality of educational programs and services, a long-standing interest of state government, has been to examine such things as the quality of buildings and facilities, the credentials of professional personnel and per-pupil expenditures.

The passage of the Elementary and Secondary Education Act of 1965, with its requirement that schools objectively assess the effects on student achievement produced by federally funded programs for the educationally deprived, focused attention on measuring the performance of students to assess the effectiveness of the schools.

The essential merit of this approach has become increasingly evident to educational decision makers at the state level, and laws mandating statewide assessment of the quality of education have been passed in many states.

Because of time constraints, inadequate budgets, and the speed with which many assessments have been mandated, state departments of education often find they lack a workable plan for assessment or the personnel to conduct it.

Assessment Principles

Experience with statewide assessment programs has led us to the formulation of the following principles, which are designed as a guide for state department personnel and others to assist them in optimizing the chances for a successful assessment.

Involve the community. Effective educational assessment demands the recognition and involvement of the entire community--legislators, educators, parents, students, business managers, labor leaders and other concerned groups.

One method of involving them is to have representatives from each group assist in determining what the goals for education ought to be. Since each group may have different priorities, this could be a time-consuming activity. The time will be well spent, however, since in addition to determining the goals the participants should also become aware of the needs and constraints of the others. For example, the legislator wants to know about how much pupil learning and development the money he appropriates for education is buying. He also must answer to his constituents, who may not reelect him if they feel he is not concerned about the quality of the education their children are getting. Since the goals for education most directly concern the students, they should have representation in deciding what those goals ought to be. Parents want assurance that their children are receiving the kind of education that will enable them to cope with the ever increasing complexity of the world in which they live.

Teachers also have an interest in assessment. Some may have negative attitudes because they feel they personally will be evaluated. In addition to the valuable contribution they can make, they will be less apt to feel threatened because they have been given the opportunity to participate in the developmental phases of the program.

The early involvement of the various interest groups should facilitate understanding and cooperation when the assessment is conducted.

Specify and define goals. After the goals have been determined, they must be defined operationally and behaviorally so they can be measured. The community should continue to be consulted in this phase--especially the educators.

An example of this type of definition is the goal "To appreciate human endeavor in the arts." One aspect of this goal would be to appreciate music. An appreciation of music could be defined in behavioral terms as the number of times tapes and records are used. This definition corresponds to the receiving and responding levels of the affective domain (Krathwohl, et al, 1964). The behavioral objective could then be measured by a frequency count of the tapes and records used in the library and those taken out for off-campus listening. The number of usages and the proportion of students involved would be an indicator of the student body's appreciation of music.

Measuring devices must have face and content validity. The instruments should contain an adequate sampling of the specified universe of content. In addition, they should be face valid. That is, the layman must be able to look at the tests and see the relationships between them and the goals being measured. If the objective is to measure understanding and the instrument contains items that are purely factual in content, the instrument would not have content validity, although it might appear to be face valid. Adequate assessment devices must present both.

Take noncognitive effects of school into account. Society is delegating more and more responsibility to the schools for developing learning outcomes which are not skills centered. The appreciation of music goal mentioned earlier is one example. Another is the development of a positive self-concept. Although these noncognitive areas are admittedly more difficult to measure, in an assessment program they must not be ignored in the early phase or they most likely will continue to be neglected as the program is enlarged.

Data presentations should be designed for lay understanding. Possibly the most crucial aspect of a successful assessment program is the reporting of results. The reports should be in terms that are understandable to the layman. Interpretation of statistical data, particularly that which requires qualification, such as test scores, is most effective when interaction between the receiver and the presenter is possible. However, there is likely to be little interaction if the results are reported in sophisticated technical terms. Four possible alternatives for use in the presentation of data are: expectancy tables based on previous year's performance; comparison with state norms; percentage of reponse to each option of key items; and description of the distribution of student scores in terms of the kinds of problems they are solving successfully and the kinds which are presenting difficulty.

Assessment must not be an end in itself. The last principle, which perhaps should have been first, is that assessment must be clearly identified as one component of the total education process. Evaluation data are collected to meet specific needs, and if the data are not related to these purposes they are useless. Assessment must provide feedback to enable decision makers at various levels to make program modifications necessary for educational improvement.

A Model for Beginning State Assessment

The politics of assessment frequently limit the number of methods available to achieve the principles which were set forth in the preceding section. As a result, state department personnel may have to work under any or all of the following constraints.

First, time schedules--especially when limited by legislative action--most often do not allow an adequate and thorough development of assessment procedures.

Second, the resources made available are usually far less than required. Thus, one must expect that compromises will be made in the operations of the program.

Third, the unavailability of adequate professional staff further complicates effective implementation of assessment activities.

Fourth, the conceptions of assessment held by the several publics who are concerned with it are frequently ambiguous and overly optimistic.

Within the context of the constraints identified, a simplified model with limited objectives can be implemented. The first objective is to collect data which will provide a status report on education within the state in specified areas of greatest interest. This model is designed to provide statewide data, not individual or school building data. The most commonly specified areas are reading and math because these basic skills are fundamental to most educational activities. It is recommended, however, that even the first-level model include data collection in at least one noncognitive area. This recommendation is made because of the human tendency to concentrate efforts upon the areas being evaluated. Therefore, the failure to evaluate noncognitive areas has the effect of focusing the educational process on the skill development segment of education to the neglect of the equally important but more difficult to measure noncognitive areas.

One such area related to the two major cognitive areas is student attitude toward school or learning. This attitude has apparent value to those concerned with the educational process and seems a reasonable area of interest in which to begin initial efforts in noncognitive data collection.

The second objective of this first-level model is to introduce operational concepts of assessment to the interested parties. These include school personnel, both administrative and teaching, state government personnel, including legislators and executive office staff, and concerned community groups. The consideration of methods of statewide educational data collection by these groups will provide them with insight into the limited nature of the kind of information available from a basic model and the problems related to obtaining it.

The third objective is to provide a plan which will enable state education personnel to gain experience in dealing with assessment problems. Among these problems are: selecting and securing data collection devices; negotiating the needs for assessment data presented by a heterogeneous public; developing communication strategies which minimize destructive conflict and optimize data utilization; and establishing the organizational structure to carry out assessment programs.

The fourth objective is to provide a method of analyzing the data to illustrate the variability of performance due to the individual differences among students and to the social context in which they live. The reason for using this method of analysis is to clarify ambiguous perceptions of the interested publics concerning these correlates of performance. Most people recognize that there are individual differences and that these differences contribute to the difficulty of learning. On the other hand, much of the public appears to expect that some simple method of instruction, if properly applied, will overcome such difficulties. The data produced by this model can and should be used to analyze contextual learning difficulties and differences, thereby requiring a major consideration of them.

Components. The components of the first-level assessment model include introductory activities, data collection materials, analysis procedures, and reporting strategies.

Introductory activities are informative and communicative in purpose. An integral part of these activities is the selection of an Advisory Committee to consider the objectives of the assessment and plans for its accomplishment.

The selection of the Advisory Committee should be given very careful thought. It should be cross sectional in nature, representing the several publics who are concerned with assessment. An additional qualification for members should be sincere interest in education as a community responsibility. It is probably not possible under the conditions for which this model was designed to convene the Advisory Committee for the several sessions which would result in optimum consideration and support of the assessment program. Therefore, the committee function would be that of a review group. Department staff, augmented as necessary by independent consultants, should prepare tentative objectives and plans for committee review. The committee's recommendations should be carefully considered, incorporated if possible, and always given the courtesy of a response.

Following the preparation of initial plans and review by the Advisory Committee, a series of regional conferences should be conducted. The conferees should represent both the community and professional educators. Special emphasis should be placed upon the effects of assessment activities in the school, and open discussion of this matter should be given substantial time on the conference agenda.

The arrangements for each conference should include provision for small group sessions to facilitate an interchange of ideas among the representatives of the different groups. The care with which the Advisory Committee was selected will be reflected in the success of these conferences because, to the degree that the participants feel they were represented in the earlier planning, they will be inclined to respond favorably to the plans.

The success of these introductory activities also depends on the content of the plans presented. The components which delineate the desired content are described next.

The major component is the measurement package. For the objective of this first-level model, presently existing devices should be selected where possible. Standardized achievement tests or minor modifications of such tests are appropriate for reading and math. The exercises used by national assessment, which are placed in the public domain upon release, are also candidates for utilization. For example, reading and math exercises are available in standardized tests, released NAEP items, and the Delaware and Michigan state assessment programs. The Delaware, Michigan, and Pennsylvania assessment programs have also used instruments that measure attitude toward learning.

In order to clearly communicate what will be accomplished with the measurement package, it is advisable to provide a content reference which will show how scores may be reported in specific skills- or behavior-related form. Content reference as used here means an example or description of the behavior implied by the response to the question. For example, the steps to provide such a reference for a sample reading comprehension item based upon paragraphs are as follows:

- Identify a series of skills tested by the specific questions asked such as making inferences, detecting mood, or recognizing factual detail.
- Prepare or select an equivalent paragraph to that presented in the test and show how it will be used to illustrate the obtained results.
- After testing, report the percentage of students who successfully respond to each kind of question, indicating accomplishment of the skill of interest.

In the case of attitudinal questions, descriptions of specific approach or avoidance behaviors evoked by situations or persons are the content reference. A Likert-type response format to this sort of item will provide information about the proportion of students who respond in each direction (approach or avoid) and also about the intensity of the responses to the behavior in question. After testing, if it is decided not to report the actual questions, the underlying constructs can be described, and the response frequencies can be related to them.

The final ingredient of the measurement component for this model is the collection of relevant educational context indicators which may be used to classify the school score distributions and thereby provide information useful in generating hypotheses about the antecedents of student performance. Examples of these indicators are socioeconomic status and teacher verbal ability (see Coleman, et al).

Data Collection. Preparing for the actual collection of data for this model requires a decision about the reference population. It may be a population of schools or individual students. If the school is the unit for analysis, it is frequently easy to secure a sampling frame because most state departments maintain a list of operating schools in their states. The problem of obtaining adequate representation within the schools, however, is not so easily solved. Assuming a high degree of variability within a school, rather large student samples are required to provide a school score with a reasonable degree of precision. Therefore, it is recommended that for this model a sample of randomly selected schools be used to generate the state educational status data specified in the first objective. An example of this procedure follows.

Suppose that there are 3,000 elementary schools in the state and that the measure of interest has a standard deviation of fifteen. In this situation, a sample of 200 elementary schools will provide an estimate of the state average which will deviate from the actual average no more than two score points with 95% confidence. If information is available which is known to relate to the student characteristic of interest, stratification of the sample will result in greater precision, and a reduction of sample size becomes possible (see Kish).

It is possible, however, to select a sample which allows analysis of individual student data although the complexity and therefore the risk of administration error increases.

The design requires that the probability of selecting any one student remain equal even though we do not know the names of all students in the state. An example of this procedure is furnished in Appendix A.

The school sample described above does not permit the analysis of any individual student output data unless some generally untenable assumptions are made about the random assignment of students to schools. Neither does such a sample permit estimates to be made for individual school districts. Because of the anticipated substantial variance among schools within districts, it would probably be necessary to sample most or all of the schools in each district in order to obtain suitable district estimates. As in the case of data collection at the state level, the model at the district level would not necessarily require the testing of all students within the selected schools. However, such all-inclusive data collection might be more feasible than within-school sampling of individuals.

District estimates would, of course, enable districts to compare themselves to other districts "like them" as defined by a wide variety of situational variables. Reports could be generated to facilitate these comparisons in terms of both district averages and district score distributions.*

Analysis Procedures. The objective of this first-level model is to provide a description of the position of the schools in the state with reference to the educational objectives deemed important in that state. The data may therefore be analyzed according to straightforward descriptive information. Frequency distributions for the state, the measure of central tendency probably expressed as a mean and a measure of variability such as the standard deviation can readily be produced. The second objective of the model, however, is to direct attention toward the different levels of achievement of student groups as a way of highlighting both differential difficulty of learning and the areas in which hypothesis generation might be productive.

For these purposes, analysis of the data should minimally include distributions classified by the qualitative information collected to reflect the educational context. This information does not have to be quantifiable, although the idea of levels may be appropriate, in order to be useful in the analysis. A minimum four-cell classification of distributions is recommended so that possible interactions may be detected. For example, the data might look like this:

*These district assessment ideas were suggested by Donald Trismen, Educational Studies.

Attitude Toward School

	High Staff Training	Low Staff Training
High SES	47.8 49.5 46.2 52.1 $\overline{X} = 50.2$ 51.0 51.7 54.7	44.6 47.4 42.6 46.8 $\overline{X} = 45.8$
Low SES	50.4 47.8 51.2 48.8 $\overline{X} = 49.5$	50.2 47.3 52.0 47.6 $\overline{X} = 48.6$ 54.5 46.8 44.3

An inspection of these data suggests that, in terms of attitude toward school, teacher training has a more significant association than socioeconomic level. There are appropriate statistical techniques for determining the probability that the observed differences are actual rather than due to chance. Two possible procedures are two-way factorial analysis of variance or a Friedman two-way analysis for ranked data.

The analysis described here is useful in identifying interrelationships which should be examined further. The purpose of this additional examination is to discover what experiences children have which may be modified to produce desired changes in output--in this case improving attitude toward learning. Observation of schools located in high-scoring and low-scoring groups in the classification tables should suggest productive ways of changing the learning situation.

The limitations of this analysis, however, include the possibility that less distinct relationships may not be revealed. Also, it becomes extremely complex to examine the effects of several conditions taken as a group.

If the data collected on both student output and the conditions of learning are quantifiable, and if there is reason to believe that relationships are fairly uniform across the range of scores from high to low, it is appropriate to use multiple correlation techniques for the analysis. These procedures allow more complex relationships to be considered and provide a method for examining the unique contribution of many variables in a systematic way. Partial and semipartial correlation techniques are included in this classification (see Nunnally).

The principal use of these correlation techniques--as in the case of the two-way classification analysis--is to identify variables which should be studied for possible influence upon the experiences students have and therefore upon what they learn. Results obtained from these technqiues do not suggest corrective action directly but are the first part of a two phase process of educational change. The second phase requires alteration of learning conditions or of the arrangement of learning experiences which can then be evaluated by a subsequent assessment.

Reporting Results. Interpretation of statistical data, particularly that which requires qualification, such as test scores, is most effective in a context in which interaction between the receiver and the presenter is possible. Therefore, the ideal method of interpretation includes a personal interface between the concerned school personnel and a presenter who knows both the nature of the data and the method of analysis. If the whole process is to be cost effective, a discussion of implications consistent with the results and suggestions of alternative courses of action must be included.

In reporting the results of the state status assessment to legislators, state boards of education, the governor's staff and other decision makers, the personal interface is extremely important. If standardized tests or NAEP exercises and procedures have been used, the results can be compared with regional and national data.

If district or school building data have been collected, more detailed methods of reporting results are needed. Several options for reporting results are provided in Appendix B.

Inter-District Comparisons

In the foregoing discussion, the primary focus is on an assessment program that views the state in its entirety (i.e., the sample is drawn from the state population for the grade level(s) of interest). Considerable interest is being directed toward the use of student performance data gathered in an assessment to compare school districts.

The following is a brief discussion of two considerations that should be taken into account when inter-district comparisons are to be made. One relates to the sample design and the data analysis. The other pertains to socio-economic and other condition variables--both situational and individual--and their criticality when inter-district comparisons are to be made.

Sample Design and Data Analysis. As was indicated earlier, some generally untenable assumptions must be made if district estimates are made from a sample of students where the population frame is the state in its entirety. Should inter-district comparisons be desired, the population from which schools or students are drawn should be the individual district. Since most school districts either maintain master lists of students or can obtain a list with minimum difficulty, the less complex method of drawing a sample is to use a simple random sample of students. It is possible, but more difficult, however, to use a two-step cluster sample technique in some districts (i.e., in step 1 select a sample of schools; in step 2 draw a sample of students from each of the previously selected schools) while using a simple random sample procedure in others. This combined approach might be feasible when exceptionally large school districts do not maintain or have available student lists. In large districts the two-step cluster sample technique might be used, while in smaller districts where student lists are available a simple random sample of students could be drawn. The major difficulties in using the two-step cluster sample technique relate to bias in local scores caused by the possibility of greater student homogeneity (i.e., less variability) in an individual school.

Most states have a wide range in the number of students enrolled in their school districts. When drawing a student sample, larger districts may be collecting data on, for example, 10% of the population, while in smaller districts it may be necessary to gather data from the entire population. When aggregating district data upon which to make inferences statewide, arithmetic weighting of district data is used when computing estimates of parameters and their standard errors. The disproportionate allocation of students in individual districts is compensated for by using inverse weights in the statistics. That is, in the larger district where 10% of the student population is sampled and therefore under-represented, the data are weighted up, while in a smaller district where data are gathered on all of the students and therefore over-represented, the data are weighted down. The correct weight for a district can be determined by $\frac{Nj}{N}$ where Nj is the number of students in the population from the jth district and N is the number of students in the total population.

Criticality of Condition Variables. When district comparisons are desired, special attention should be directed toward those conditions of learning that may be associated with student performance.

Student background characteristics such as socioeconomic status, attitudes and aspirations have been found to be associated with (not to be confused with caused by) student achievement. Furthermore, the other school variables such as the quality of the instructional staff (e.g., staff training) and the availability of financial resources also have shown an association with student achievement.

These variables become critical when inter-district comparisons are to be made. For example, comparing an inner-city school district serving children from economically deprived families with the affluent suburbs surrounding could be grossly misleading. The inner-city school district, when the environment of its students is considered, might be making a more substantial contribution to student performance than its more affluent neighboring school districts, when the environment of their students is taken into consideration. Other examples could be cited; however, the important point is that as nearly as is possible school districts should be compared with those that have similar characteristics that would be difficult to change by educational policy decisions (e.g., socioeconomic status).

In comparing similar districts, it is important that categorization of districts reflect those difficult-to-change variables that are associated with differences in output. Although the ultimate objective of assessment is to provide information which will enable decision makers to improve the educational performance of all children, it is naive to expect that such improvement will result immediately. Therefore, the condition variables should be considered in designing alternative school programs which show promise of improving student performance (e.g., greater utilization of prior student experiences). Statistical procedures for determining categories include expectancy tables, regression analysis, analysis of variance, and analysis of covariance.

Summary

This plan provides some suggestions and ideas for initiating a statewide educational assessment program.

Guiding principles for an assessment should be: 1) to specify and define educational goals in terms of measurable outcomes; 2) to involve various publics extensively; 3) to use measurement instruments having face and content validity; 4) to include noncognitive student behaviors; 5) to present the results in a form understandable by those outside the professional education community; and, finally, 6) to view assessment not as an end unto itself but as a means of providing useful information to decision makers.

The objectives of the initial statewide assessment program should be: 1) to collect student performance data that can provide a status report on the quality of education in those goal areas identified as having high priority; 2) to introduce the concept of assessment and its usefulness as a source of information for both decision makers and concerned parents or taxpayers; 3) to provide a starting point whereby those managing the statewide effort may gain useful experience in operating the program; 4) to develop a method of data analysis that can illustrate the variability of performance due to individual differences among students and to the social context in which they live.

The essential components of the initial statewide assessment should include such introductory activities as the provision for an advisory committee and for selecting its members and conducting meetings to inform interested citizens and school personnel of the nature, scope, and methodology of the assessment program. Other essential components are data collection materials, data analysis procedures, and reporting strategies.

In order to make inter-district comparisons, the sample must be representative of the individual districts rather than the entire state. Simple random sampling of students in the districts is recommended when feasible. The more complex two-step cluster sample technique may be used in large districts as an alternative.

When comparisons are made, the condition variables must be considered in categorizing the districts, or the comparisons will not have meaning. Expectancy tables, regression analysis, analysis of variance-and analysis of covariance are statistical procedures that may be used to determine categories.

REFERENCES

Coleman, J. S., et al. Equality of educational opportunity. Washington, D.C.: U. S. Government Printing Office, 1966.

Kerlinger, F. N. Foundations of behavioral research. New York: Holt, Rinehart and Winston, Inc., 1964.

Kish, L. Survey sampling. New York: J. Wiley, 1965.

Krathwohl, D. R., Bloom, B. S., & Masia, B. B. Taxonomy of educational objectives, handbook II: affective domain. New York: David McKay Company, 1956.

Nunnally, J. C. Educational measurement and evaluation. New York: McGraw-Hill 1964.

APPENDIX A

Sample selection - individual student sampling unit - equal probability

a. Arrange the schools in any order, with the number of students in the grade of interest specified for each school.

b. Assign sequential blocks of numbers to each school. The size of the block is determined by the number of students in the target grade. Inflate the block size by a specified percentage to allow for increases in enrollment. This percentage should be taken into account when determining sample size so that the desired sample size will remain after the shrinkage caused by empty sampling frames (i.e., numbers in the block not having a corresponding individual).

c. Using a suitable random selection procedure (table of random numbers, computer program, etc.), determine the numbers which are to be included in the sample.

d. Assign selected numbers to each school block, converting them to the sequential number within the block (e.g., in the school with block 2998-3092, 2998 becomes 1, 2999 becomes 2, 3024 becomes 27).

e. Instruct school coordinators to arrange class lists in any order and assign the numbers 1 through N to the total group of students (not recycling through classes).

f. After the assignments have been made, the instructions are to open the envelope containing the randomly selected numbers and administer the test to those students whose names on the school list correspond to these numbers.

APPENDIX B

Options for Reporting Results

Option one - interface:

Step 1

Process data and prepare written reports.

A. Alternative formats

1. Expectancy tables based on previous year's performance
2. Comparison with state norms
3. Percentage of responses to each option of "key" items--those where content is face valid and which relate most highly to other items in the subscale, thereby reflecting well the concept measured
4. Description of the distribution of children in terms of the kinds of problems they are successfully solving and the kinds which are presenting difficulty; reporting of available response patterns for high scorers, middle scorers, and low scorers

Step 2

Assemble and train a corps of data interpreters

A. Sources of personnel

1. State department personnel
2. College and university staff, graduate students, interns, and so forth
3. School personnel
4. Intermediate unit office staff

No single source can provide enough personnel to comprise the necessary teams, but a combination may make possible sufficient numbers to provide one interpretive visit to each district.

B. Training

1. Sessions concerning test development, item to goal correspondence, reliability, group statistics versus individual statistics, and presentation of comparative data classified by school variables (Methodology includes presentations using actual state data.)

2. Sessions exploring implications of results and identifying strategies for the development of alternative action plans where need is indicated.

These sessions should include specialists in the areas assessed. The objective of these sessions is to stimulate action through provision of suggested approaches or avenue of exploration based on assessment results.

Step 3

A. Presentation sessions scheduled with each school district

It is recommended that a work group comprising administrative staff, teachers, parent advisory representatives, at least one board member, and other interested people be included.

Step 4

Organize a consortium of educators across the state who can serve as a resource to the school personnel who will be working on altered approaches to learning and personal development for their students.

The resource requirement for a program of this sort is dependent in part upon voluntary professional commitments of the state's educators. It also requires allocation of funds for personnel time on the part of many institutions in addition to the Department of Education. If properly approached, there is enough professional commitment among professional educators to achieve some degree of this resource allocation.

For a state with 500 districts, the requirements of the interpretation team are approximately as follows. In a one-month period, the 500 districts would be visited on the basis of one and one-half days per district. Allowing some surplus for contingencies, 78 persons working 11 days each will be required. Based on a per diem expense of $25 plus an average transportation cost of ten cents per mile for 150 miles, the total cost will be less than $30,000.

Such an approach to the problem of reporting and utilizing assessment data has the best chance of not only assuring positive follow up but also of reducing hostility in both the schools and their communities.

Option two - regional interface model:

It is recognized, however, that resource constraints may not allow so comprehensive a commitment. Therefore, a less optimistic alternative can be conceived.

This model would follow the same basic steps outlined in Option One, but rather than providing individual district data presentations as outlined in Step 3 it would utilize a series of regional meetings in which about two representatives from each district would participate. The format suggested for the regional meetings is the model used for the training sessions designed for the data interpreters (Step 2 of Option One). The corps of data interpreters in this case would be smaller since 30 districts can be accommodated in each regional meeting requiring only four data interpreters for each session. The person days required are thus reduced to about 120. If we assume that local districts will pay transportation and housing costs of their representatives and also assume the professional investment of the state's educators, the cost for this kind of program would be about $6,000. It must be recognized that the technical experts used for training data interpreters will not be available for the 17 regional meetings and that their contribution would be filtered through two "student" levels, the data interpreter and the school district regional meeting attendees. However, the concentrated thinking about the problems and the utilization of assessment results that such a program would encourage would still stand a fair chance of having an impact on the education in a state.

Option three - lecture-discussion model:

In order of desirability, a third level of data presentation might be to conduct five regional meetings with an average of 200 participants comprising two representatives from each of one hundred districts. Real but nonidentifiable state data could be presented in a visual lecture-discussion mode in an auditorium setting with questions limited to a small sub-group of school representatives. These interrogators could collect their questions from discussion sections prior to the question and answer period. Such meetings could be handled by State Department personnel augmented by some representation from the corps of experts available within the state. Following these regional meetings, the assessment data would be released to each local school district--preferably scheduled to allow time for district staff consideration before it becomes public.

Option four - mail-out kits:

The least desirable, least expensive, and not only least useful but possibly useless presentation is a mail-out kit of charts, lists, and printed discussions which would be sent to the superintendent in each district. To optimize its very slim chance for utility and positive impact, the mail-out kit should include the following components.

First, the data should reflect the maximum amount of knowledge on the part of the State Department of Education about the school district to which it is directed. Second, it should be personalized in a form which describes what the students are like who represent several points on the distribution of data for the school.

For example, a student in the upper quarter of the score range on the reading scale could be described in terms of skill content of the question to which he responds correctly. The actual pattern of responses should be included. The report could read "Students in the highest quarter of the score range tend to answer correctly items which require inference two out of three times and items which require locating factual detail four out of five times. There are 178 (20%) students in this range from your school. For comparison, students from other schools of your community type (Type 3) respond correctly to inference items three out of four times and to factual detail questions also three out of four times. Seventeen percent of all community Type 3 students score in the highest quarter."

A similar discussion of middle and low scores should be provided. Backup material describing community type and resource availability should also be included. In addition, any of the written alternate forms of presentation suggested in Option One as material for the more desirable personal interpretation could be provided. This published form of presentation could also be used as a supplement to any of the options.

In summary, the most productive interpretive format for state assessment data is conceived to be an individually tailored personal presentation to school district personnel. The least productive interpretation is a written report. A state department of education is urged to locate the resources to do personalized data interpretation rather than a less involved and therefore less expensive mail-out.